UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Nano-Electronique et Nano-Technologies**

Arrêté ministériel : 7 août 2006

Présentée par

« Mohamad Hairol / JABBAR »

Thèse dirigée par **« Dominique / Houzet »**
codirigée par **« Omar / Hammami »**

préparée au sein du **Laboratoire GIPSA-Lab, Grenoble and ENSTA ParisTech, Paris**
dans **l'École Doctorale EEATS**

Méthodologies de Conception ASIC Pour des Systèmes sur Puce 3D Hétérogènes à Base de Réseaux sur Puce 3D

Thèse soutenue publiquement le **« 21/03/2013 »**,
devant le jury composé de :

M. Ian O Connor
Professeur, Ecole Centrale de Lyon, France, Président, Rapporteur
M. Paul Franzon
Professeur, North Carolina State University, États-Unis, Rapporteur
M. Said Hamdioui
Professeur, Delft University of Technology, Pays-Bas, Membre
M. Kholdoun Torki
Directeur Technique, Circuits Multi-Projets (CMP), France, Membre
M. Yvain Thonnart
Ingénieur de Recherche, CEA-Leti, France, Membre

Université Joseph Fourier / Université Pierre Mendès France /
Université Stendhal / Université de Savoie / Grenoble INP

Title in English:

ASIC Design Methodologies for 3D NoC-Based 3D Heterogeneous MPSoC

Keywords:

3D IC, Exploration, MPSoC, NoC, Physical design implementation

Titre en Français:

Méthodologies de Conception ASIC Pour des Systèmes sur Puce 3D Hétérogènes à Base de Réseaux sur Puce 3D

Mots clés:

3D IC, Exploration, MPSoC, NoC, Mise en œuvre de conception physique

ABSTRACT

For many years, Moore's Law has been the primary driving force enabling the evolution of semiconductor industry with the ability to double the transistor count on a silicon die for every two years. However, shrinking transistor dimensions, also known as CMOS scaling to be able to design and manufacture higher performance devices has become much more difficult than it is previously as we are approaching very deep submicron technologies such as 20 nm and beyond. The issues of design complexity and the exponential increase of cost to manufacture devices based on these very deep submicron technologies are among the great hurdles currently being faced by the industry making it unattractive performance per cost solution. The transition to 450 mm (18") wafer to help reducing manufacturing cost for advanced process technology and the development of extreme ultraviolet (EUV) lithography tools are also facing technical difficulties that remain to be solved in the next several years while at the same time requires multi billion dollar investment to build new manufacturing facilities as well as new processing equipments.

3D integration has been around since decades ago but only now the industry is paying great attention to this technology as a result of economical and technical difficulties that arise from the transistor shrinking in 2D technology. It has been the subject of extensive research in the industry and academia due to benefits it could potentially offer such as higher performance, lower power consumption, larger memory bandwidth, small form factor and support for heterogeneous technology integration making it suitable for several application domains particularly in mobile devices. 3D technology could provide higher memory bandwidth through its excessive vertical connections using TSV or microbumps as in wide I/O memory architecture and can also accomodate high memory capacity when using memory-on-logic or memory-on-memory stacking. Shorter vertical interconnection between stacked dies or wafers as well as reduction of horizontal wirelength due to stacking will eventually provide higher performance per watt. However, there are also some challenges that exist in 3D technology and they have to be solved before it can be widely adopted as a mainstream technology for high volume production such as high temperature effect, testing of 3D architecture and most importantly for the designers is 3D design tools, specifically the tools that are capable of doing 3D synthesis, 3D place and route as well as 3D optimization at each step.

With the recent trend of mainstream multiprocessor technology that is moving towards increasing the number of processing cores to support higher performance applications, Network-on-Chip (NoC) has become the primary technology in meeting the demand of high performance, scalability

and flexibility for processor's and Intellectual Property (IP) cores' communication. Works on multiprocessor and NoC architecture in 3D technology have been carried out for many years covering various issues such as partitioning method and NoC topologies but most of the prior works only consider software simulation for the performance analysis where the results is less accurate and therefore cannot be truly used for evaluating the benefits bring by 3D technology. The need for performance analysis from design implementation results is highly desirable to be able to make the right conclusions regarding the potential benefits it offers. In this thesis, we study the 3D NoC architectures through physical design implementations using real 3D technology being implemented in the industry. Based on the routed netlists, we conduct performance analysis to evaluate the benefit of 3D architecture compared with its 2D implementation. But firstly, we present our initial work designing and implementing a 2D NoC-based MPSoC architecture on FPGA intended to identify design issues related to the 2D MPSoC design.

Based on the proposed 3D design flow focusing on timing verification by leveraging the benefit of negligible delay of microbumps structure for vertical connections, we have conducted partitioning techniques for 3D NoC-based MPSoC architecture including homogeneous and heterogeneous stacking using Tezzaron 3D IC technlogy. Design and implementation trade-off in both partitioning methods is investigated to have better insight about 3D architecture so that it can be exploited for optimal performance. Using homogeneous 3D stacking approach, NoC architectures are explored to identify the best topology between 2D and 3D topology for 3D MPSoC implementation. The architectural explorations have also considered different process technologies highlighting the wire delay effect to the 3D architecture performance especially for interconnect-dominated design. Additionally, we performed heterogeneous 3D stacking of NoC-based MPSoC implementation with GALS style approach and presented several physical designs related analyses regarding 3D MPSoC design and implementation using 2D EDA tools.

Finally we conducted an exploration of 2D EDA tool on different 3D architecture to evaluate the impact of 2D EDA tools on the 3D architecture performance. Since there is no commercialize 3D design tool until now, the experiment is important on the basis that designing 3D architecture using 2D EDA tools does not have a strong and direct impact to the 3D architecture performance mainly because the tools is dedicated for 2D architecture design. Integrating manual tools (scripts to constraint the design) to the 2D EDA tools to design 3D architecture is the common method to achieve performance benefit but this method loses the most important design step of 3D optimization that normally exists in the 2D EDA tools when designing 2D architecture.

RÉSUMÉ

Pendant de nombreuses années, la loi de Moore a été la principale force motrice permettant l'évolution de l'industrie des semiconducteurs avec la possibilité de doubler le nombre de transistors sur une puce de silicium pour tous les deux ans. Toutefois, la diminution des dimensions des transistors, appelés aussi mise à l'échelle CMOS pour être en mesure de concevoir et de fabriquer des appareils plus performants est devenu beaucoup plus difficile qu'elle ne l'est déjà que nous nous approchons de technologies submicroniques profondes tels que 20 nm et au-delà. Les questions de la complexité de la conception et de l'augmentation du coût exponentiel pour la fabrication de dispositifs basés sur ces technologies submicroniques profondes sont parmi les grands obstacles actuellement rencontrés par l'industrie qui rend la performance peu attrayante pour solution économique. La transition à plaquette de 450 mm (18 ") pour aider à réduire le coût de fabrication pour la technologie avancée des processus et le développement d'outils de lithographie par ultraviolets extrême (EUV) sont également confrontés à des difficultés techniques qui restent à résoudre dans les prochaines années alors que dans le même temps nécessite investissements de plusieurs milliards de dollars pour construire de nouvelles usines ainsi que des équipements de traitement de nouvelles.

L'intégration 3D a été autour depuis des décennies auparavant, mais que maintenant l'industrie accorde une grande attention à cette technologie en raison des difficultés économiques et techniques proviennent du contraction du transistor dans la technologie 2D. Il a fait l'objet de recherches approfondies dans l'industrie et le milieu universitaire en raison de certains avantages qu'elle pourrait potentiellement offrir telles de meilleures performances, faible consommation d'énergie, la mémoire de bande passante plus large, plus petit facteur de forme et de soutien pour l'intégration des technologies hétérogènes ce qui convient pour plusieurs domaines d'application particulièrement dans des appareils mobiles. La technologie 3D peut fournir une mémoire de bande passante plus élevée par l'intermédiaire de ses connexions verticales excessives à l'aide TSV ou microbilles selon l'architecture de la mémoire éventail d'E/S et peut également accueillir une grande capacité de mémoire pour l'utilisation de la mémoire-à-mémoire logique ou mémoire-à-logique d'empilement. Interconnexion verticale plus courte entre des matrices empilées ou de plaquettes ainsi que la réduction des longueur de fil horizontale due à l'empilement finira par offrir de meilleures performances par watt. Cependant, il y a aussi des défis existent dans la technologie 3D, et ils doivent être résolus avant de pouvoir largement adopté comme une technologie majeure pour la production de volume élevé, comme l'effet de la température élevée, les tests d'architecture 3D et le plus important pour les concepteurs sont des outils de conception 3D, spécifiquement les outils

qui sont capable de faire la synthèse 3D, le lieu et l'itinéraire en 3D ainsi que l'optimisation 3D à chaque étape.

Avec la tendance récente, la technologie multiprocesseurse déplace vers l'augmentation du nombre de cœurs de traitement pour supporter les applications haute performance, réseau sur une puce (NoC) est devenue la principale technologie pour répondre à la demande des performances, une évolutivité et une flexibilité élevées pour le processeur et propriété intellectuelle (PI) de communication de cœurs ».Travaux de multiprocesseur et de l'architecture NoC dans la technologie 3D ont été réalisées depuis de nombreuses années sur divers sujets tels que la méthode de partage de topologies NoC mais la plupart des travaux antérieurs ne considère pas le logiciel de simulation pour l'analyse de la performance où les résultats sont moins précis et ne peuvent donc pas être vraiment utilisées pour évaluer les bénéfices apporter par la technologie 3D. La nécessité d'une analyse de la performance des résultats de la mise en œuvre de conception est fortement souhaitable d'être en mesure de faire les bonnes conclusions quant aux avantages potentiels qu'ils peuvent offrir. Dans cette thèse, nous étudions les architectures 3D NoC grâce à des implémentations de conception physiques en utilisant la technologie 3D réel mis en œuvre dans l'industrie. Sur la base des listes d'interconnexions en déroute, nous procédons à l'analyse des performances d'évaluer le bénéfice de l'architecture 3D par rapport à sa mise en œuvre 2D. Mais tout d'abord, nous vous présentons notre travail initial de conception et la mise en œuvre d'un NoC 2D basé sur l'architecture MPSoC sur FPGA pour but d'identifier les problèmes de conception liés à la conception 2D MPSoC.

Sur la base du flux de conception 3D proposé en se concentrant sur la vérification temporelle en s'appuyant sur l'intérêt de retard négligeable de la structure de microbilles pour les connexions verticales, nous avons mené des techniques de partitionnement de NoC 3D basé sur l'architecture MPSoC y compris empilement homogène et hétérogène en utilisant la technologie Tezzaron 3D IC. La conception et la mise en œuvre de compromis dans les deux méthodes de partitionnement est étudiée pour avoir un meilleur aperçu sur l'architecture 3D de sorte qu'il peut être exploitée pour des performances optimales. En utilisant l'approche empilage 3D homogène, les topologies NoC est explorées afin d'identifier la meilleure topologie entre la topologie 2D et 3D pour la mise en œuvre MPSoC 3D sous l'hypothèse que les chemins critiques est fondée sur les liens inter-routeur. Les explorations architecturales ont également examiné les différentes technologies de traitement en mettant en évidence l'effet de la technologie des procédés à la performance d'architecture 3D en particulier pour l'interconnexion dominant du conception. En outre, nous avons effectué l'empilage 3D hétérogène pour la mise en œuvre MPSoC avec l'approche de modéle de GALS et présenté

plusieurs analyses de conception physiques lié concernant la conception 3D et la mise en œuvre MPSoC en utilisant des outils de CAO 2D.

Enfin, nous avons procédé à une exploration de l'espace de conception d'architecture en 3D en utilisant l'outil du lieu et d'itinéraire 2D pour but d'évaluer l'impact de l'utilisation des outils de CAO 2D sur la performance d'architecture 3D. Comme il n'y a pas de l'outil de conception 3D jusqu'à présent dans le marché, l'expérience est importante pour le motif que la conception de l'architecture 3D en utilisant les outils EDA 2D n'a pas un impact fort et direct pour la performance d'architecture 3D principalement parce que les outils est dédié à la conception de l'architecture 2D. L'utilisation d'outils de support aux outils de CAO 2D pour concevoir l'architecture 3D est une méthode courante pour obtenir le gain de performances, mais cette méthode perd l'étape de conception le plus important de l'optimisation 3D existent normalement dans les outils de CAO 2D lors de la conception d'architecture 2D.

DEDICATION

To my wife, Siti Noor Saadah and my family.

ACKNOWLEDGMENTS

First and foremost, I would like to thank Allah for always helping and giving me courage throughout the period of this study to complete the research works as well as writing this thesis.

Thank you to Professor Omar Hammami and Professor Dominique Houzet for being my thesis directors during this PhD study. Without them, I will not be involved in this fascinating technology doing state of the art research works where not many people have the opportunity as I have. I really appreciate all the time they have spent discussing and evaluating the research works and also for the technical as well as non-technical advices they have given to improve research outcomes as well as to teach me to be a better researcher in the future. I pray for them to have long lasting happiness in whatever they do and to success in whatever dream they want to achieve.

I am really honoured to have Professor Ian O Connor of EC Lyon and Professor Paul D. Franzon of North Carolina State University as my PhD thesis reviewers. Their comments have greatly helped me to improve the quality of this manuscript. I am grateful to have Professor Said Hamdioui of Delft University of Technology, Dr. Kholdoun Torki of CMP Grenoble and Mr. Yvain Thornnart of CEA-Leti for being in my PhD defense committee.

I would like to thank Dr. Kholdoun Torki of CMP for the help on ASIC design and most importantly on Tezzaron Design Kit. Thank you also to Dr. Alejandro Chagoya who has helped me with the EDA tools at CIME NANOTECH, MINATECH to run all the experiments in this thesis.

For my colleagues at ENSTA as well as at GIPSA LAB, specifically to Dr. Abir M'zah, Khawla Hamwi and Dr. Xinyu Li, thank you for helping me in various difficult times during my stay here either in academic as well as in life. Thank you for the friendship and I hope we will meet again.

I would like also to express deep gratitude to my sponsor both University Tun Hussein Onn Malaysia (UTHM) and Ministry of Higher Education of Malaysia (MOHE) for funding my PhD study.

Finally, my thank you goes to my wife who is very patient while I was spending most of the time doing research works as well as writing papers and thesis. Thank you also to my parents who always pray for my success in academic from far away.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

VERSION FRANÇAISE

Introduction

Conceptions électroniques ont connu une croissance rapide au cours des dernières années et qui s'est déclenchés par l'introduction des smartphones et des tablettes dans les marchés. Petit facteur de forme, de meilleures performances et moins d'énergie figurent parmi les exigences d'appareils mobiles afin de fournir plus petit, moins cher, plus rapide appareils électroniques au grand public. La Figure 1 montre l'évolution du nombre d'éléments de traitement dans les appareils grand public portables SoC selon l'International Technology Roadmap semi-conducteurs (ITRS). Comme la montre la figure, dans un proche avenir, le nombre d'éléments de traitement devrait augmenter de plus de 100 processeurs. En outre, la taille de la mémoire est également prévue d'augmenter considérablement à l'avenir avec l'augmentation du nombre d'éléments de traitement.

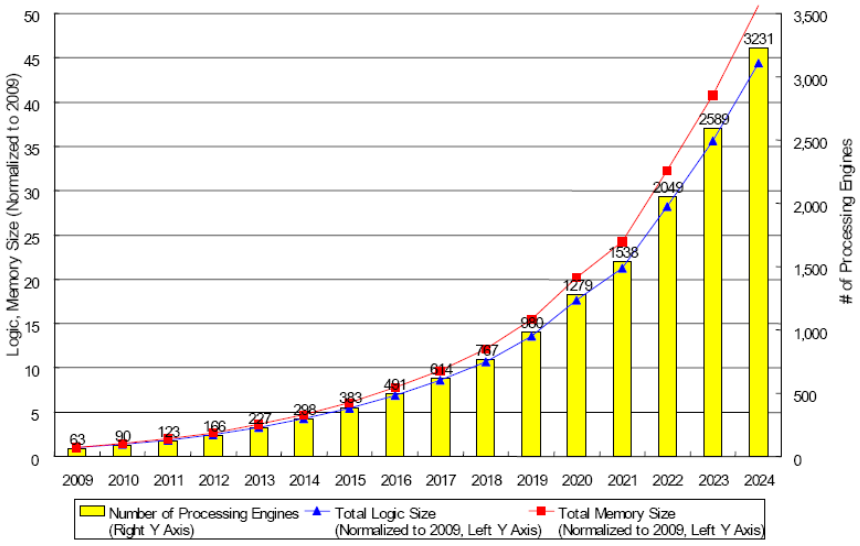


Figure 1: Nombre de processeurs et l'évolution de logique de mémoire de taille dans les systèmes électroniques

Améliorer la performance de la conception de processeur unique grâce à la fréquence d'horloge plus élevée pose inconvénient de forte consommation et de l'architecture multiprocesseur qui a été mise en place lorsqu'un dessin ou modèle à plusieurs processeurs simples fonctionnent à une fréquence plus basse et l'approvisionnement en basse tension. Nous passons d'une vaste architecture de calcul d'architecture de communication vaste place de cette architecture multiprocesseur. Cependant, la conception de haute performance architecture multi cœur du processeur nécessite plusieurs défis à résoudre, comme différents interfaçage de IP Core, automatisation de la conception, vérification et

la programmation logicielle [1]. Afin de répondre à la demande de l'exigence de communication, réseau sur puce (NoC) est développé surmonter la limitation de l'architecture de bus base telles que celles à long temps de retard en raison de la politique d'arbitrage, le câblage complexe et donc d'augmenter la consommation d'énergie du système. En contraste, NoC offre une évolutivité en augmentant le nombre de processeurs.

Au départ, nous nous appuyons sur la fonction de mise à l'échelle CMOS pour avoir plus de performance qui est obtenue par réduction de la dimension physique du transistor afin que les transistors de nombreux autres peuvent être emballés dans une seule puce et ainsi augmenter les performances grâce à une architecture pipeline plus profondément. Toutefois, le passage vers les nœuds de processus plus petites introduit beaucoup de grands défis économiques et technologiques et en même temps diminuer les avantages de performance à tous les nœuds d'échelle [2]. En outre, la limitation de la mise à l'échelle CMOS comme les limites maximales de tension et de la variabilité dispositif ont également eu des répercussions techniques de conception au niveau système et circuit où les techniques de conception supplémentaires sont nécessaires pour permettre l'amélioration des performances et de réduction de la puissance croissante de la réduction des coûts [3].

D'autre part, il existe une alternative pour augmenter les performances de la conception électronique sans passer par un chemin difficile avec mise à l'échelle CMOS appelé l'intégration 3D. Cette technologie permet la construction de circuits en 3 dimensions (3D) des structures en empilant les plaquettes en plusieurs couches. Cette nouvelle technologie offre des avantages potentiels de la plus grande vitesse, faible consommation d'énergie, l'intégration des technologies hétérogènes, petit facteur de forme et de haute densité d'intégration appareil. Contrairement à la technologie CMOS mise à l'échelle, l'intégration 3D est une solution prometteuse pour conduire l'avenir de circuits VLSI pour soutenir la demande continue des systèmes électroniques de haute performance. Ces avantages seront décrits en détail dans les sections suivantes.

L'intégration 3D, la longueur d'interconnexion à long fil métallique est réduite à la racine carrée de la longueur de l'intégration 3D grâce à la connexion en utilisant soit verticale courte TSV ou microbumps. Ceci améliore la vitesse où il réduit le retard RC du long fil d’interconnexion dans l'architecture 2D et également de réduire le nombre de tampons le long du fil d'interconnexion dans laquelle finalement la consommation d'énergie est réduite aussi. Intégration 3D supporte également l'intégration de technologies telles que hétérogène numérique, analogique, RF et la technologie MEMS, où ils peuvent être fabriqués selon la leur technologie de processus optimal et ensuite

empilées avec l'autre technologie des procédés. Malgré l'avantage amené par cette technologie, il est également confronté à plusieurs obstacles critiques tels que les problèmes thermiques, la structure de livraison de puissance et de synthèse d'horloge en arbre. Comme l'intégration 3D empilées plusieurs couches de silicon actifs, la densité du dispositif par unité de volume est augmenté et augmente donc la densité de puissance. Un autre effet est qu'il existe des points chauds thermiques dus au différent profil de consommation de puissance de blocs logiques dans chaque couche. Cela crée gradient thermique sur la puce qui crée des variations qui pourraient affecter l'intégrité et la fiabilité des dispositifs au fil du temps. Bien que certaines des techniques d'essai pour l'architecture 2D peut être étendu pour permettre de tester l'architecture 3D tels intégré d'autotest intégré (BIST) et la méthode de boundary scan, mais la fabrication de TSV introduit nouveau défaut du mécanisme tels que des shorts ou ouvre due à un mauvais alignement et micro-voids et donc nécessite une nouvelle approche pour tester ces défauts. Voici quelques principales questions en matière de technologie d'intégration 3D qui doivent être surmontés avant de pouvoir être commercialement viable dans de nombreux produits.

Motivations De Recherche

Comme il s'agit d'une technologie relativement nouvelle à l'heure actuelle, de nombreuses questions sont encore à l'étude approfondie de l'industrie et du milieu universitaire. En regardant la publication dans des conférences et des revues, nous avons réalisé qu'il y a un manque de vraie conception et la mise en œuvre effectuée pour avoir des résultats plus réalistes sur l'analyse de la performance de cette technologie. En particulier, les travaux de recherche en 3D NoC architecture ont été fait principalement à l'aide de logiciels de simulation basé sur le cycle-précis simulateur qui fournissent des résultats irréalistes qui sont insuffisantes pour évaluer les avantages et les inconvénients de la technologie 3D. Par conséquent, l'une des principales motivations de ce travail est de procéder à l'analyse des performances à l'aide de la conception et de la mise en œuvre réelle de la technologie 3D grâce à la technologie qui sont disponibles à utiliser. Grâce à cette approche, nous cherchons à avoir des résultats plus réalistes et donc pourrait nous aider à mieux comprendre les offres de compromis pour utilisent cette technologie. A part cela, nous concentrerons notre travail sur l'aspect architectural de cette technologie que nous avons utilisée spécifiquement la technologie 3D de Tezzaron à deux niveaux pour l'aspect technologique. La plupart des dispositifs électroniques actuels ont plus d'un cœur de traitement afin d'avoir une plus grande capacité pour exécuter des différentes applications avec des performances supérieures. Cela est dû au fait que les performances de l'architecture monoprocesseur ne peut pas être encore améliorée à cause de le mur de puissance et de le mur de mémoire. Par conséquent, il est intéressant de savoir comment cette

technologie 3D pourrait être utilisée pour surmonter les problèmes multiprocesseurs qui se posent aujourd'hui pour être en mesure d'améliorer ses performances. Plusieurs travaux ont été fait effectuer la conception et la mise en œuvre de l'architecture multiprocesseur utilisant la technologie 3D où un certain nombre de travaux ont montré poignée amélioration des performances tout à fait significative en utilisant la technologie 3D quand comparer avec l'architecture 2D. Cependant, aucun des précèdent travaux d'analyse de l'architecture des performances de 3D NoC qui est basé sur les résultats de conception et de mise en œuvre, ce qui est l'objectif principal de ce travail. Par ailleurs, l'enquête de topologies de NoC en architecture 3D basé sur la conception et la mise en œuvre des résultats aussi n'ont pas été réalisées. Il est intéressant de comprendre quel genre de topologies de NoC (2D ou 3D) est meilleur dans l'architecture 3D en termes de performances où nous ne l'avons pas considérer comme lors de la conception d'une architecture 2D.

Contributions

1. Proposition de méthodologie de conception originale pour la conception d'architecture en 3D à l'aide des outils de CAO 2D disponibles principalement axés sur la vérification temporelle 3D. Cette vérification temporelle 3D est possible parce que la connexion verticale entre les niveaux est créée en utilisant microbumps qui a retard négligeable dans cette technologie particulière 3D. La méthode de conception proposée 3D a été utilisé pour les expériences dans cette thèse menée pour étudier les différentes mis en œuvre de l'architecture qui sont réalisables en utilisant la technologie 3D.
2. NoC présenté des topologies d'exploration dans l'architecture 3D grâce à la mise en œuvre de la conception physique motivé par les études antérieures dans la littérature qui a effectué l'analyse des performances de la mise en œuvre du logiciel. Nous avons conçu et mis en œuvre en 3D mesh NoC topologie à l'aide du routeur 3D pour la connexion verticale entre les niveaux et 3D maille empilés NoC topologie à l'aide de l'architecture d'empilage homogène routeur 2D en 2 niveaux d'architecture 3D basé sur la technologie Tezzaron et comparé ses performances avec l'architecture 2D. Enfin, nous avons proposé une nouvelle topologie de NoC d'architecture 3D qui est la topologie hexagonale qui offre de meilleures performances que d'autres topologies dans la mise en œuvre de la technologie 3D en raison des liens entre l'égalité de fil-routeur.
3. La mise en œuvre d'empilage hétérogène 3D de GALS architecture multiprocesseur par l'empilement de l'architecture NoC sur le dessus du processeur en raison du nombre limité d'œuvres de l'architecture 3D basé sur la mise en œuvre GALS. Dans cette étude, nous avons analysé les performances de l'architecture hétérogène empilées en 3D qui ont été manuellement la partition en 2 niveaux et par rapport à son architecture correspondance 2D à la conception

identifié.

4. Ayons effectué une exploration de l'espace de conception de l'architecture MPSoC 3D pour analyser l'impact des outils de CAO 2D pour sa performance comme la qualité de calendrier, la consommation d'énergie et mesure de longueur. Depuis, il est compréhensible de la limitation de l'utilisation des outils de CAO 2D pour concevoir et mettre en œuvre l'architecture 3D, cette étude a examiné l'impact sur les performances de l'architecture 3D 2D lorsque des options d'outils EDA, en particulier le placement et le routage des options est variée nous permet de comprendre plusieurs questions de mise en œuvre importants qui n'ont pas été souligné précédemment. Nous nous concentrons sur le calendrier et les options d'optimisation d'énergie dans l'outil 2D CAO électronique pour l'exploration parce que les deux mesures sont parmi les paramètres les plus essentiels qui sont considérés lors de la conception de l'architecture 3D.

Technologie 3D: Vue D'ensemble, Les Avantages Et Les Défis

L'intégration 3D offre méthode moins difficile de parvenir à une intégration plus élevé pour les besoins de l'application transistor actuelle par rapport au transistor d'échelle pour les nœuds technologiques plus petits. Avec les technologies de semi-conducteurs d'aujourd'hui, la technologie 3D est possible d'être conçus et mis en œuvre à un coût relativement faible. En utilisant la technologie 3D TSV est une approche où les matrices ou de plaquettes sont empilées et en utilisant TSV pour leur interconnexions électriques dans un seul paquet. Cette technologie offre une meilleure densité des interconnexions entre les matrices et les plus petits en raison de la structure de raccordement vertical situé à l'intérieur de la zone de matrice. Il existe aussi une autre forme de ce type de technologie connue sous le nom 2.5D à l'aide de plusieurs matrices où est placé au-dessus d'un élément d'interposition de silicon (interposition active ou passive), qui se compose de plusieurs couches de métal d'interconnexion formée avec TSV que la liaison entre matrices et de l’interface externe.

L'un des principaux avantages de la technologie 3D, c'est que la longueur d'interconnexion à long fil est réduite en raison de l'empilement. Il peut réduire les cent micromètres de fil d'interconnexion horizontales de l'architecture 2D à une longueur quelques micromètres en utilisant la structure TSV. Ceci améliore la vitesse où il réduit le retard RC du fil d'interconnexion et également de réduire le nombre de tampons le long du fil d'interconnexion et éventuellement le retard global. Une expérience sur l'architecture 3D FFT [4] et d'une architecture de microprocesseur [5] a montré l'amélioration de la vitesse de conception 3D. Pour l'architecture 2D, le nombre de couches métalliques augmente de manière à faire face à l'augmentation du nombre de transistors dans la

matrice de silicon qui jusqu'à 12 couches de métal pour la technologie de pointe. En ce qui concerne l'intégration 3D, la réduction des fils d'interconnexion globaux due à la réduction de la congestion de routage éventuellement augmente les performances. Stratégie de partitionnement a une forte influence à l'amélioration de la latence et son évolutivité lors de l'empilement des couches supplémentaires [6].

Le retard de TSV devrait être considéré lors de la mesure d'amélioration de ses performances. Les résultats expérimentaux montrent que le retard de TSV est compris entre 35 ps à 135 ps [7] et 16 ps pour 20 µm de hauteur, ce qui est inférieur au fil de retard dans l'architecture 2D par exemple 219 ps de 2500 µm longueur de fil de 4,5 GHz de vitesse. Le r de TSV est en fonction de plusieurs facteurs tels que le diamètre, la hauteur, le pitch et la technologie. La hauteur et le pitch du TSV largement affecter son retard alors que le diamètre TSV a un effet de petite taille [8]. La résistance TSV a moins à se prononcer à son effet retard de la capacité TSV [9]. En termes de technologie, basée sur la technologie SOI-3D aura moins de retard TSV que le vrac de la technologie CMOS en raison de la dimension plus petite de TSV qui réduisent le retard de RC [6]. Comme nous l'avons inclure plus de couches dans la structure 3D, supérieur retard de TSV sera remarqué due à l'augmentation de la numération TSV.

Avec l'existant de l'interconnexion verticale comme nous empilons des plaquettes, ce qui permet plus de possibilités d'optimisation de la conception qui ne peut être fait qu'en utilisant l'architecture 2D. Par exemple, nous pouvons avoir grande variété d'optimisation du partitionnement dans différents niveaux d'empilement de mémoires et les processeurs afin d'optimiser les performances de communication [4]. En dehors de cela, en utilisant l'outil de partitionnement automatisé pour la conception en architecture 3D pourraient fournit également l'amélioration des performances considérables en contraste avec la méthode le partitionnement manuel optimisé.

La consommation d'énergie peut également être réduite en raison de la réduction de la longueur du câble d'interconnexion qui diminue la capacité du câble [10] et de réduire le nombre de répéteurs. En outre, l'intégration 3D est non seulement dépassé en termes de performances, il est également évolutive que la conception deviennent plus complexes, comme par exemple l'amélioration de la consommation d'énergie d'environ 11%, 21% et 46% pour les 12, 36 et 72 bits Kogge-Stone adder [11].

En raison de la forte densité d'interconnexion entre des matrices empilées ou de plaquettes, d'architecture 3D peut être utilisé pour atténuer les problèmes de mémoire mur en fournissant

courtes liaisons verticales et permet également une plus grande capacité de mémoire sur puce nécessaire en particulier par la grande architecture multicore [12]. L'architecture wide I/O est une autre approche pour atténuer les problèmes de mur de mémoire en offrant des données d'une bande passante élevée grâce à plus grand nombre de broches d'I/O pour l'accès mémoire [13].

Construire l'architecture empilement 3D permet petit facteur de forme de l'architecture 2D. L'épaisseur de la puce après l'empilement est de plusieurs centaines de microns, ce qui est relativement faible par rapport à la puce 2D classique. L'avantage de petit facteur de forme permet une intégration à haute densité. Par exemple, l'empreinte puce est réduite de 44% pour les quatre couches empiler contre deux couches de la pile de 65 nm [14].

L'intégration des technologies hétérogènes est moins complexe que dans la conception 2D où les architectures différentes, par exemple analogiques, RF, capteur, mémoire pour être intégré sans processus de fabrication difficile, car chaque architecture qui est produit en utilisant leur propre technologie de processus optimale et puis ils sont intégrés dans la structure 3D à l'aide méthodes telles que collage de plaques. En outre, l'architecture hétérogène peut également être implémenté en utilisant la technologie des procédés différents, tels que 95 nm pour le processeur à 65 nm de mémoire tel que démontré dans [15]. Ceci permet aux applications SoC avec une meilleure capacité de répondre aux besoins des systèmes embarqués tels que le traitement en temps réel et une faible consommation électrique et aussi un support pour la conception de SoC avenir qui est une structure très hétérogène [16]. Du point de vue SoC qui a des blocs numériques et analogiques dans une matrice, l'intégration 3D aussi surmonter le problème d'isolation du bruit de l'architecture de signal mixte 2D parce analogique / RF et des composants numériques peuvent être placés séparément dans différentes couches de silicon [17].

TSV est une méthode qui utilise par l'intermédiaire de l'autre côté les différentes couches de silicon actif. Utilisations pour TSV est Tungstène (W) [18], le Cuivre (Cu) [19] [20] et Poly-Silicon (Poly-Si) [21]. Poly-Si matériau est stable et a moins d'effet sur la caractéristique du dispositif que les autres matériaux. Cependant, Cuivre ou Tungstène est plus approprié pour la cause de diminution de la résistance TSV. Cuivre est le plus couramment utilisé car il a une bonne conductivité thermique par rapport à Tungstène et Poly-Si. Cependant, comme on le verra plus tard dans les défis et des enjeux de 3D, Cuivre TSV créer un effet de stress dû à la grande différence de coefficient de dilatation thermique (CTE) entre le substrat de silicon et de cuivre, ce qui n'est pas le cas pour Tungstène TSV. Une comparaison détaillée de la via filling material se trouve dans [22]. Tungstène a plus retarder par rapport à Cuivre TSV pour n'importe quelle taille de diamètre et est

donc utilisé dans la recherche [23]. TSV peut être formée soit à l'aide du MEIR ou perçage laser qui, comme représenté dans la Figure 2 affectant sa taille. Utilisation MEIR, aussi connu comme Procès Bosch, est une méthode largement utilisé qui peut produire un rapport d'aspect élevé, mais au détriment de la hausse du coût par rapport à la méthode de forage au laser. Utilisation la méthode de perçage au laser limite le diamètre TSV à environ 10 µm. En outre, cette méthode est un processus de série et ne donc pas adapté pour des conceptions de nombre élevé de TSV [24].

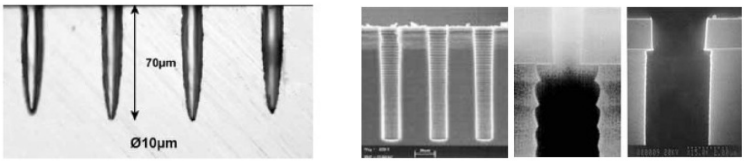


Figure 2: Formation TSV en utilisant (a) forage au laser et (b) DRIE

Il y a différentes techniques pour mettre en œuvre TSV comme par l'intermédiaire du premier, par l'intermédiaire du milieu ou par l'intermédiaire de dernier, tel qu'illustré à la Figure 3. Dans première technique, TSV est formée avant que la structure BEOL. Par conséquent, nous avons une taille relativement petite par rapport aux deux autres méthodes [25]. Alors que par le dernière approche TSV se forme après la formation BEOL, ce qui entraîne plus grande dimension TSV. La dernière méthode est un processus plus difficile car la formation de TSV pourrait endommager les appareils qui ont déjà été formés. Le processus de fabrication TSV se compose de plusieurs étapes qui sont de forage, d'isolation, de remplissage ou de métallisation, la formation FEOL, la formation BEOL, la manipulation d'attachement, plaquette amincissement et le traitement au verso. L'ordre dépend des techniques de formation TSV soit par l'intermédiaire dernière ou par l'intermédiaire premier. Comme la manipulation de fine plaquette est un grand défi, elle peut être évitée par l'amincissement de la plaquette après collage avec une autre plaquette d'épaisseur. Il pourrait également empêche une baisse de rendement en raison des processus supplémentaires pour le collage de la poignée de plaquettes et de décollage.

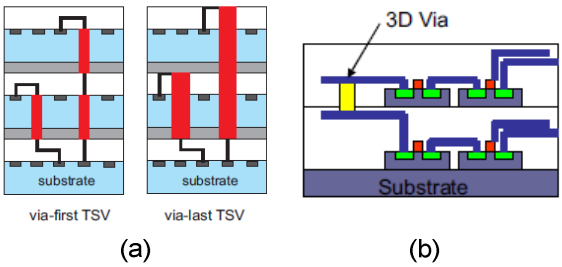


Figure 3: TSV méthodes d'empilage (A) par-premier-dernier et par l'intermédiaire d'CMOS en nombre (b) par l'intermédiaire de premier TSV-CMOS SOI dans [26]

Méthodes d'empilement peut être implémenté de plusieurs manières telles que wafer-to-wafer, die-to-wafer ou die-to-die. La méthode de plaquette à plaquette est surtout utilisée pour l'intégration 3D en raison du faible coût que les deux autres méthodes. Cependant, il souffre de faible rendement en raison de die de mauvais rendement d'adhérence par rapport à d'autres méthodes de collage qui prennent en charge known good die (KGD). Un autre inconvénient d'empilage plaquette à plaquette est qu'elle est limitée aux die de même taille dans les plaquettes rendant offre un débit de production élevé. La méthode de die-to-die engendre de coût élevé dû à la liaison de chaque filière, mais peut être utilisé pour différentes tailles de die obligataires.

Du point d'orientation de collage, il existe en plusieurs méthodes, comme face-to-face, face-to-back et back-to-back comme le montre la Figure 4. Pour 2 niveaux comme dans la mise en œuvre technologie 3D Tezzaron, l'orientation face-to-face est la meilleure façon où l'inter-moule connexions microbumps est utilisée, et donc ne bloque pas les couches de routage. Pour plus de 2 niveaux comme dans MIT Lincoln Lab 3 niveaux de technologie, à la fois face-to-face et face-to-back l'orientation est utilisé lorsque toutes les connexions inter-tier se fait à travers la structure TSV.

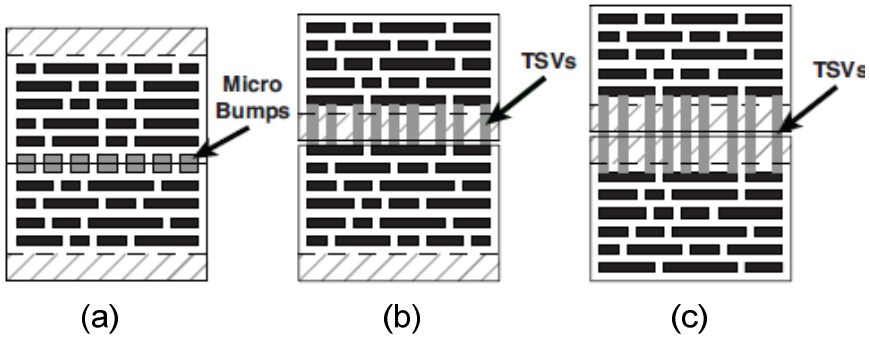


Figure 4 : Orientations d'empilement (a) *face-to-face* (b) en *face-to-back* (c) *back-to-back*

En termes de mise en œuvre lien physique, plusieurs méthodes peuvent être employées comme du métal à la liaison métallique, l'oxyde de collage direct et collage adhésif. En règle générale, la liaison métal-a-métal est meilleur car elle donne une connexion mécanique et électrique entre plaquette et Cu est le matériau le plus couramment utilisé. Cependant, il souffre d'un traitement à température élevée, par exemple supérieure à 350ºC. L'alignement de collage est le paramètre clé pour atteindre la densité d'interconnexion haute de la liaison métal-a-métal [27]. Cette température élevée doit être soigneusement surveillé car il peut endommager la couche inférieure et affecter l'appareil. Collage adhésif utilise un traitement à basse température. Cependant, il y a des possibilités de contamination par le matériau adhésif sur les appareils. Parmi les matériaux utilisés

sont benzocyclobutène (BCB), polyamide et pyralène. BCB est le plus couramment utilisé car il a la plus grande force de liaison qui est supérieure à 20 MPa [28]. Le collage direct utilise substrat d'oxyde de silicon ou de la matière de collage [29]. Il se fait à température ambiante, puis recuit à haute température pour obtenir covalente Si-O-Si. Par conséquent, il a la plus grande force de collage par rapport à d'autres méthodes. Le problème est qu'il est très sensible à la contamination par exemple 1 µm de diamètre des particules pourrait créer 1 cm de diamètre vide lors du collage de huit plaquettes [28]. Techniques de collage hybrides a également été rapportée en utilisant Cu avec collage [30].

Tezzaron 3D Technologie

Tezzaron technologie 3D est basé sur le niveau plaquette d'empilage. La tranche est collée au moyen d'une liaison en utilisant un matériau métallique Cu thermique tel que représenté sur la Figure 6 [31]. Tezzaron a été mis au point plusieurs architectures TSV, l'un d'eux est la technologie FaStack. Ils atteignent une précision d'alignement pour la tranche d'environ 0,5 µm. Tezzaron utilisé des méthodes via le premier face-à-face de liaison et ses propriétés d'empilement sont indiquées dans Figure 7. Plusieurs puces de test 3D a été démontrée en utilisant cette technologie comme un capteur CMOS, 3D FPGA, ASIC à signaux mixtes et processeur / mémoire pile. Parce que la plaquette est amincie après le collage, il n'ya donc pas besoin d'un processus de manipulation des plaquettes aider à réduire les pertes de rendement en raison des processus supplémentaire de liaison et de dé liaison de la poignée plaquette.

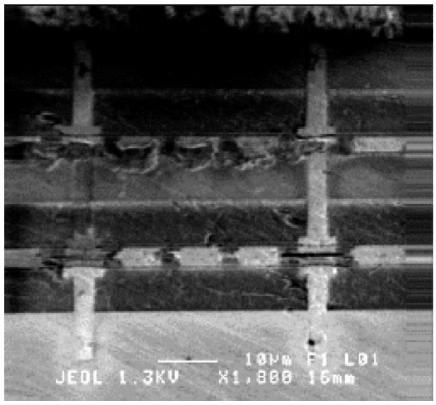


Figure 6: Tezzaron face-to-face superposition à l'aide microbumps

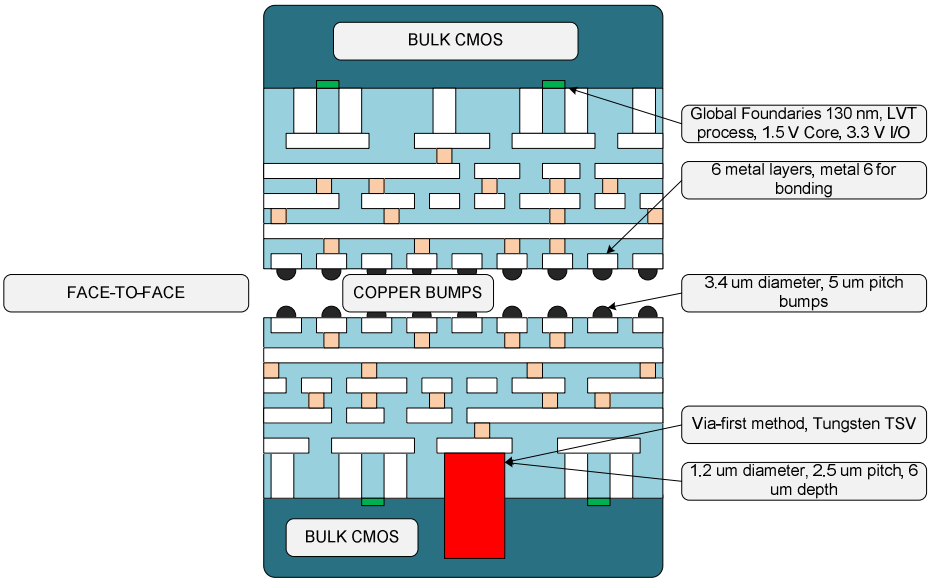


Figure 7: les paramètres de la technologie Tezzaron 3D

Malgré les avantages qu'il offre, il existe plusieurs problèmes qui doivent être résolus pour faire de l'intégration 3D peut être appliquée dans les appareils électroniques grand public. La nature de la 3D stacking provoqué une surchauffe ne peut pas facilement être transférée hors de la puce, en particulier la chaleur générée loin du réservoir de chaleur. Contrairement à l'architecture 2D laquelle la chaleur générée pour tous les composants peut être transférées directement sur le heat sink par dissipateur thermique car il peut être placé juste au-dessus des composants. L'importance de l'effet d'empilement dans la structure 3D est augmentation de la pic de température [32] [33] dans la puce où il peut s'élever à plus de 100 ℃. Deux choses sont très importantes à la suite de cette température élevée qui est la variation de température et point d'accès. Ces deux éléments influents sur la fiabilité de la puce sont le temps moyen de taux d'échec (MTTR) et temps de claquage (TTBD). La puissance de fuite augmente exponentiellement avec la température. Chaque augmentation de 15 ℃ de température provoquent la variation du délai d'interconnexion autour de -15% à 10%. L'augmentation de la température est également à l'origine d'électro-migration qui augmente de façon exponentielle et réduit éventuellement la durée de vie du produit par quatre fois [34]. Plusieurs méthodes ont été proposées pour les techniques de gestion thermique pour résoudre le problème thermique dans l'intégration 3D telles que l'élevage thermiques qui placent les blocs le plus souvent passer près de heat sink [35] et l'utilisation de thermal vias pour transférer la chaleur de la puce [36]. La tension thermique est un autre effet du problème thermique lors de l'intégration à l'aide TSV. Cela est dû à la propriété différente de la Silicon CTE, Cu, dioxyde de silicon et W. La tension thermique est une source de timing variation autour de ± 10% pour une cellule individuelle [37]. Thermique induite par la tension dans l'intégration 3D provoque des fissures à l'interface du substrat de silicon et TSV et Cu entre les interconnexions et low-k isolateur [38].

Question de rendement est également un autre facteur important qui doit être pris en considération pour la structure 3D. L'intégration 3D réduit le rendement global en raison du processus de fabrication qui seront devenus plus petits que d'autres filières est empilé [39]. Modèle de rendement a été développé pour aider à la prise de décision concernant le montage des compromis pour l'intégration 3D comme le nombre de piles à construire et à ce qui est la taille du die optimale [40]. Certaines des techniques pour améliorer les pertes de rendement tels que les ressources redondantes telles que les mémoires et les réseaux de capteurs et en améliorant le processus d'intégration 3D [41]. Autre défi pour l'intégration 3D est l'essai. L'essai de TSV est un problème parce que la taille de la sonde est grande (35 µm) par rapport à la petite taille TSV tels que 5 µm de diamètre avec 10 µm pitch [42]. C'est parce que normalement cette sonde est utilisée pour tester l'architecture 2D qui est normalement utilisé pour les essais de collage du fil.

Flux De Conception

La conception 3D proposée dans notre travail est montré dans la Figure 8 tire profit du petit retard des connexions inter-rang en raison de la structure de microbumps. Ce flux de conception 3D générique peut être réutilisé pour n'importe quelle architecture 3D ciblage technologie 3D Tezzaron contrairement à certains flux de conception rapporté qui est adapté à la technologie particulière. En outre, par rapport aux flux de conception précédente, nous procédons à la vérification 3D à chaque étape du flux dans backend et frontend. Pour la vérification après synthèse 3D, il est possible, car les retard des bosses est très faible et donc négligeable. Par conséquent nous pouvons avoir estimation de performance précoce de la conception 3D après l'étape de synthèse et de gagner du temps parce que nous pouvons avoir la modification architecturale pour satisfaire les spécifications de performance avant de procéder à l'étape de lieu et l'itinéraire que cela prend du temps assez long en particulier pour la conception relativement importante avec un nombre très élevé de microbumps ainsi que la fixation d'autres violations en DRC.

Le conception est d'abord divisé en 2 blocs correspondant à 2 niveaux d'architecture 3D au niveau RTL. Par la suite, avant le fin flot de conception est réalisée à l'aide Synopsys Design Compiler avec un débit de budgétisation chronogramme. Flux de synchronisation budgétaire est une méthode de distribution des contraintes temporelles entre les blocs logiques (dans notre cas, les blocs de partition) de sorte que chaque bloc peut être mis en œuvre séparément et optimisé par leur refoulement propre fin. Tout d'abord, la conception de haut niveau 3D contenant deux blocs partitionnés sont analysées et élaborées avant contraintes temporelles 3D est appliqué. Ensuite, la commande synchronisation budgétaire est exécutée pour générer des contraintes temporelles pour

chaque partition où il sera utilisé pour compiler et générer netlist de chaque partition. Une fois la compilation de blocs partitionnés est terminée, la compilation de niveau haut de conception 3D est alors effectuée et le synchronisation est analysé. En cas de violations de synchronisation existent à chaque partition ou le bloc à la conception 3D de haut niveau, les contraintes temporelles 3D modifient est en relâchant la fréquence d'horloge et le flux budgétaire est répété.

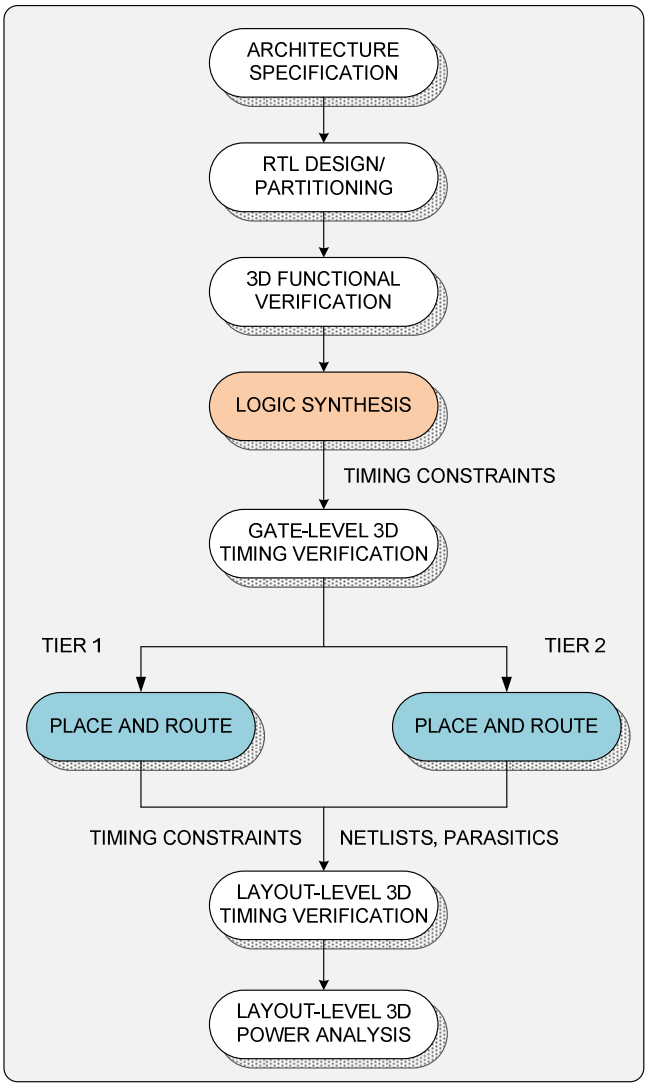


Figure 8: Flot de conception 3D mettant l'accent sur la vérification temporelle

Netlists générés à partir des niveaux de porte et des contraintes de cadencement de chaque partition, est placé et routé. Figure 8 montre l'endroit détail et de débit itinéraire où il est comme flux de conception 2D normale, sauf que l'insertion supplémentaire microbumps pour les connexions inter-rang, y compris les mesures pour des missions bosses pour inter-rang connexions. L'emplacement des microbumps est déterminé manuellement en regardant le net, il sera connecté afin de ne pas avoir à long longueur horizontale avant d'atteindre l'interconnexion verticale. Une fois l'emplacement est fixé, les tableaux de bosses sont créés, puis la numérotation des bosses est

modifié de sorte qu'il est plus facile d'attribuer aux signaux. Enfin, broches physiques sont créés dans chaque microbumps d'être en mesure d'acheminer par le NanoRoute dans SoC Encounter. Au cours de l'étape de planification de l'énergie, les microbumps est également utilisé pour la connexion à la masse entre les deux niveaux de puissance en alignant la position de bosses en tant que tel palier est recouverte sur le dessus de l'autre une fois que l'empilement est réalisé. Une vaste gamme de microbumps est formée par un fils électriques au sol étant donné que ces cours peuvent être assez fournies aux autres niveaux pour assurer un fonctionnement correct. Elle est suivie par le flux de conception 2D classique qui est mise en place, l'horloge de synthèse en arbre et de routage avec l'optimisation est guidé par les contraintes temporelles générées par des flux extrémité avant.

Exploration De Conception Physique Des Topologies 3D Noc

La nécessité d'une mise en œuvre modèle physique pour évaluer les performances 3D NoC permettrait une analyse plus précise de la quantité d'avantages pourraient apporter l'intégration 3D pour l'architecture NoC d'analyse de la performance grâce à des méthodes de simulation tel que rapporté par plusieurs ouvrages précédemment. Dans ce chapitre, nous étudions plusieurs architectures NoC 3D en tenant compte de ses propriétés physiques de conception afin d'évaluer sa vitesse et à l'amélioration du pouvoir sur son architecture 2D. Bien que l'étude précédente a démontré analytiquement que la réduction de l'espace physique de tuiles aidera à réduire le délai de réseau et de consommation d'énergie, nous avons découvert que ce n'est pas vrai pour tous les cas où de nombreux facteurs tels que les nœuds technologiques, des techniques de partitionnement et les caractéristiques des conceptions de déterminer la performance de l'architecture 3D. Les résultats montrent aussi que la topologie 2D surpasse topologie 3D dans le contexte de retard de fil à la fois pour l'architecture 2D et 3D architecture de la technologie de pointe. Cette étude contribue à la connaissance actuelle de la conception de circuits 3D intégrés par l'étude de la performance des routeurs d'empilage 3D et de l'impact de la topologie de la performance de l'architecture 3D NoC basé sur les dispositions en déroute.

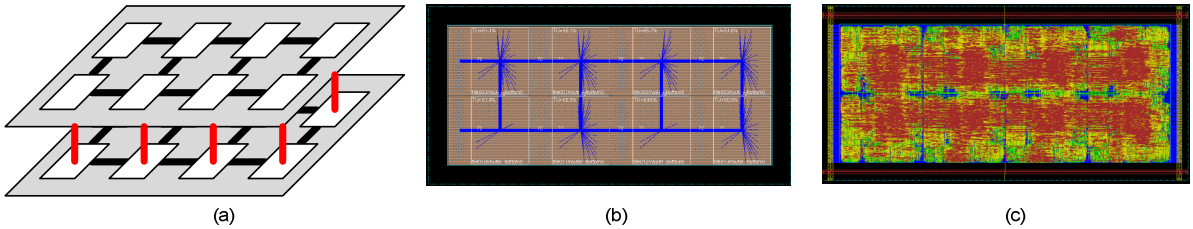
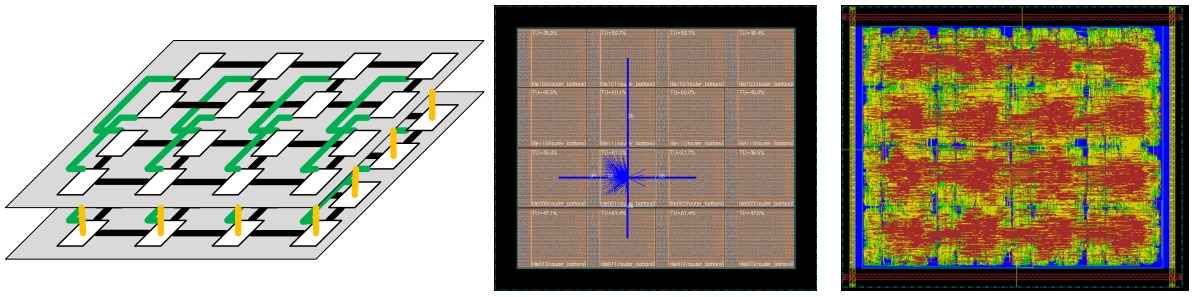
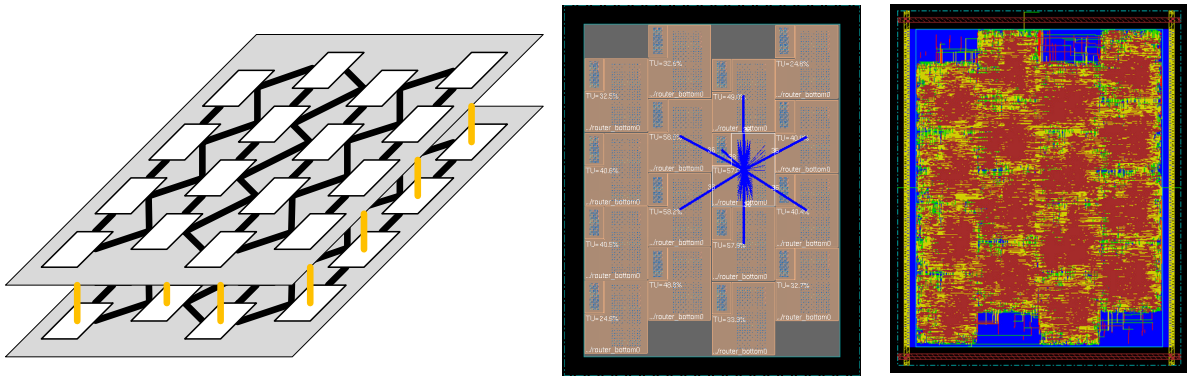


Figure 9: 3D Mesh NoC (a) diagramme (b) floorplan (c) mise en déroute



(a) (b) (c)

Figure 10: 3D Stacked Mesh NoC (a) diagramme (b) floorplan (c) mise en route



(a) (b) (c)

Figure 11: 3D Stacked Hexagonal NoC (a) diagramme (b) floorplan (c) mise en déroute

Pour la 3D Mesh l'architecture NoC, le NoC 3D est réalisé sur deux niveaux, où chaque niveau a des blocs identiques comme indiqué dans Figure 9. C'est la plus simple de l'approche 3D d'architecture NoC laquelle nous venons de prendre une copie d'une tuile (un routeur et une interface réseau) et le mettre sur une autre tuile. Cette 4x2x2 mesh NoC architecture est basée sur l'architecture 3D routeur qui a des liens verticaux pour les connexions inter-rang entre les routeurs. Ces liaisons physiques verticales indiquées dans la couleur rouge sont basée sur les liaisons logiques verticales dans chaque routeur 3D.

D'autre part, nous avons considéré une autre approche pour la construction de l'architecture NoC 3D comme le montre la Figure 10. Plutôt que d'empiler les tuiles les unes sur les autres, au contraire, nous cartographier la NoC 3D sur la mise en 2D, puis le partitionner en deux niveaux. Comme le montre la figure, les liens logiques verts représentent les connexions verticales entre les routeurs 3D tandis que les liens physiques verticales de couleur orange est fondamentalement les liens logiques 2D au sein de la structure logique de la NIU et le routeur. Cependant, cette méthode de partitionnement nécessite plus grand nombre de connexions inter-rang de 3D Mesh NoC. Plutôt que d'utiliser des outils automatiques tels que HMetis à la partition de la conception, nous nous concentrons sur la division du chemin de données en deux parties et placez-le en deux niveaux afin

de préserver les propriétés homogènes de l'architecture bloc carrelage entre les deux niveaux. Par exemple, un 4 profondeurs 32 bits Largeur du FIFO dans le routeur est divisé à 4 profondeurs de 16 bits Largeur. Pour la commande de logique, nous essayons d'équilibrer la logique dans les deux niveaux d'avoir surface de puce égale. L'inconvénient de cette structure est que les liaisons filaires entre les routeurs ne sont pas égales pour tous les routeurs en raison des liaisons filaires verticaux sont plus longues que d'autres liens.

En raison de l'inégalité des liaisons filaires entre les routeurs dans l'architecture 3D Stacked Mesh NoC en raison des liens logiques verticales (lignes vertes sur la Figure 10 (a)), afin de mieux l'optimiser, nous avons proposé une nouvelle topologie ayant la même longueur d'inter-routeur physique Liens appelée topologie hexagonale illustré à la Figure 11. Comme dans le NoC 3D Mesh empilées, les liens oranges représentent 2D fils logiques dans la structure logique de la NIU et le routeur et est utilisé pour former les connexions physiques entre les niveaux verticaux. Comme nous ne pouvons pas floorplan le carreau pour créer zone hexagonale qui a six arêtes de longueur égale utilisant l'emplacement actuel et l'outil de la route, donc nous floorplan la tuile en créant une zone rectangulaire à l'aide de l'équation (a / 2) 2 + b2 = c2, où a est la hauteur de la tuile, b est la largeur des carreaux, et c est la distance physique direct entre les deux tuiles afin de déterminer la taille de chaque tuile. Bien que cette topologie est égale longueur physique d'interconnexions de routeur, il ya des zones vides en raison de la nature de cette disposition topologie qui peut être utilisé pour la structure NoC supplémentaires, telles que les infrastructures de surveillance où la capacité de surveillance devront augmenter plus grande surface pour assurer un fonctionnement fiable pour le NoC.

Table 1: Comparaison des performances des architectures 3D NoC dans 130 nm Technologie

Parameters	3D Mesh NoC	3D Stacked Mesh NoC	3D Stacked Hexagonal NoC
Vertical connections per tier	1763	6261	7255
Nµmber of links per router	370	370	444
Core area (mm^2)	3.24	4.37	5.43
Total mesure de longueur (m)	12.48	14.01	17.03
Nµmber of gates	295,956	295,510	338,337
Longest path delay (ns)	4.20	4.60	4.73
Power consµmption @ 333 MHz (W)	1.44	1.25	1.40

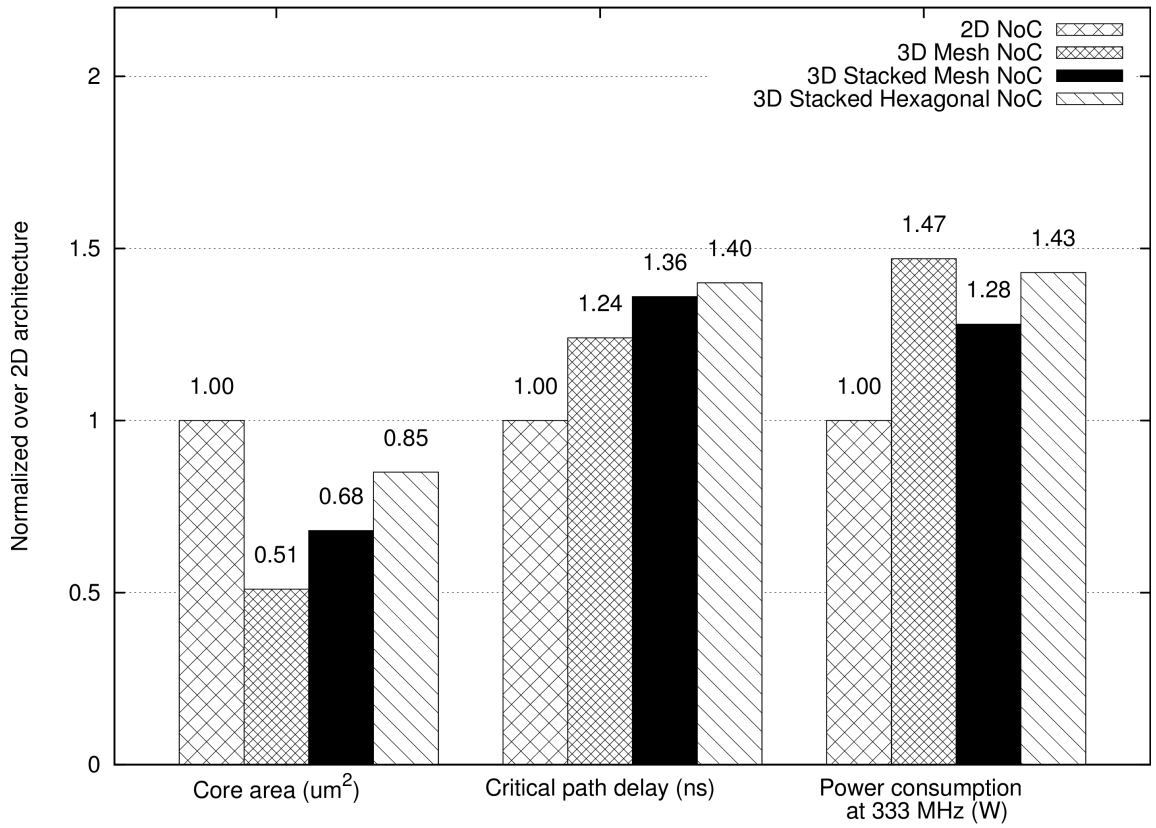


Figure 12: Comparaison des performances des architectures NoC 3D sur 2D NoC en 130 nm technologie

Table 2: Comparaison des performances des architectures 3D NoC en 45 nm technologie

Parameters	3D Mesh NoC	3D Stacked Mesh NoC	3D Stacked Hexagonal NoC
Vertical connections per tier	1763	6261	7255
Nµmber of links per router	370	370	444
Core area (mm^2)	0.79	0.91	1.01
Total mesure de longueur (m)	5.5	5.9	6.5
Nµmber of gates	255,088	257,220	290,896
Longest path delay (ns)	3.23	3.33	3.59
Power consµmption @ 333 MHz (W)	0.23	0.24	0.26

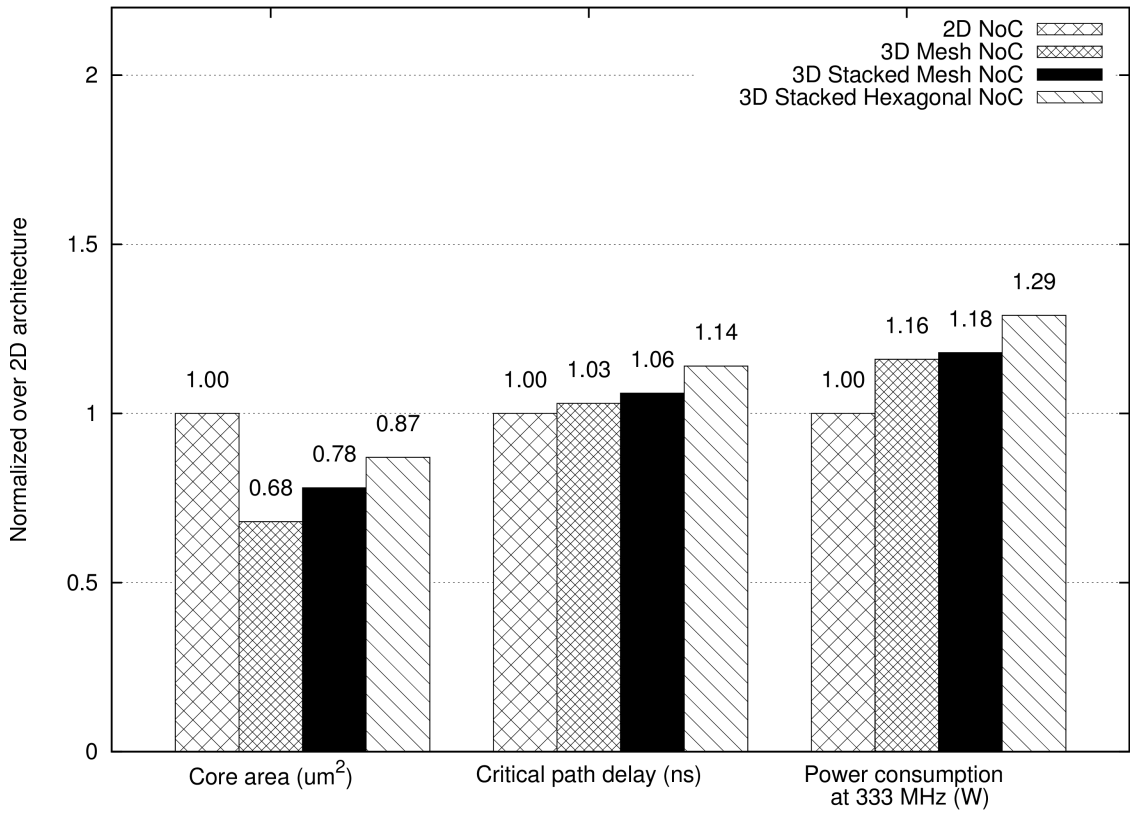


Figure 13: Comparaison des performances des architectures NoC 3D sur 2D NoC en 45 nm technologie

Pour une technologie plus ancienne, comme 130 nm et au-dessus, l'effet de la longueur du câble n'est pas significatif et le retard dans les chemins critiques est principalement déterminé par le retard des portes et des retards de fil-non. En raison de cette raison, architecture 3D offre peu ou aucun avantages de performance de contrôle sur la conception en 2D. Comme le montre la Table 1 et Figure 12, les architectures 3D NoC ne profite pas à la consommation de la vitesse et de puissance. La consommation d'énergie est plus élevée dans les architectures 3D en raison des barrières supplémentaires ainsi que mesure de longueur augmenté. Dans cette étude, nous avons utilisé simples méthode de partitionnement pour partitionner la conception 2D en 2 niveaux d'architecture 3D et les résultats montrent que la 3D n'améliore pas les performances de conception 2D. Cependant, certaines études ont montré que l'utilisation de l'outil de partitionnement automatique pourrait apporter une amélioration des performances sur l'architecture 2D en utilisant l'ancienne technologie telle que 130 nm et 180 nm [43] [44]. Le partitionnement est très important dans la conception 3D principalement pour des technologies plus anciennes.

Pour les modèles en 45 nm utilisées dans cette étude, les architectures 3D encore ne fournissent aucune amélioration par rapport à sa conception 2D comme indiqué dans Table 2 et Figure 13. Cependant, elle montre une tendance à la réduction de l'écart entre les architectures 3D et 2D

comparés avec les résultats en utilisant la technologie 130 nm nous montre dans la Figure 12. Si l'on regarde la région, nous pouvons voir que cette conception consomment très petite zone (moins de 1 μm2) et c'est la raison principale pour laquelle il n'y a aucune amélioration obtenue en utilisant la technologie 45 nm. Des travaux antérieurs ont montré des dessins de grande taille (environ 36 μm2 dans l'architecture 2D), l'amélioration substantielle des performances (réduction de 75% en retard le plus long chemin) qui pourrait être réalisé sur l'architecture 2D en utilisant la même technologie 45 nm, car mesure de longueur devient importante [45]. Table 3 montre l'extrapolation de retard fils de 22 nm basée sur la technologie du retard fil obtenir à partir d'un chemin critique dans cette étude (résultats en 45 nm) et les données du rapport d'interconnexion ITRS 2007 pour fil intermédiaire. Cette extrapolation est destiné à montrer que lorsque le modèle utilisé dans cette étude est réaliste grande, nous allons voir une amélioration pour la topologie NoC proposé hexagonale empilés architecture 3D. Du 3 mm entre routeur mesure de longueur de la 3D Mesh NoC, nous calculons la longueur du câble pour la 3D Stacked Mesh et 3D Stacked Hexagonale et en utilisant $x=\sqrt{2} \cdot a$ et $x=1{,}633 \cdot a$ respectivement un, où a est la longueur inter-routeur pour la 3D Mesh NoC et x est le nouveau routeur inter-longueur pour chaque topologie. La longueur du câble pour la 3D Mesh NoC 3D Stacked et Mesh NoC est égal parce 3D Stacked Mesh a la moitié de la superficie de Mesh 3D, mais a le double de la longueur inter-routeur pour 3D logiques liens verticaux. Comme on peut le voir dans le tableau, le délai fil est de plus en plus important de 16 nm et la technologie ainsi il engendre un fort impact sur le délai du chemin critique en particulier pour la 3D Mesh NoC et 3D Stacked Mesh NoC (en raison de logiques liens verticaux entre les routeurs) car il a plus des liaisons filaires entre les routeurs. La 2D empilés Mesh NoC surpasse la 3D Mesh NoC et 3D Stacked Mesh NoC dans le délai fil et finalement le retard total. Cependant, la 3D Stacked NoC Hexagonal se révèle être une meilleure amélioration que 2D empilés Mesh lors de l'utilisation de technologies plus petits.

Hétérogène 3D D'empilage Pour L'architecture Multiprocesseur Avec NoC

Dans ce travail, nous explorons l'architecture 3D hétérogène d'empilage et d'essayer de voir l'influence en ce qui concerne les performances de l'architecture 2D à la 3D par rapport homogène empilement présenté dans le chapitre précédent. Par ailleurs, nous avons également mener une expérience pour montrer que vertical microbumps pitch est un paramètre important à prendre en considération lors de la planification de faire de l'architecture 3D telles que le partitionnement et floorplanning bien qu'il ne inconvénients de routage blocage et grande zone de départ que en TSV. Implémentations de conception physique ont été réalisées en faisant varier des microbumps pitch pour l'interconnexion verticale de la logique et de bloc de mémoire pour étudier l'effet de

l'architecture et des performances 3D.

Table 3: L'extrapolation de retard fil intermédiaire des architectures NoC 3D utilisant des technologies différentes

Technology / NoC topology		Gate delay (ns)	Wire delay (ns)	Total delay (ns)
45 nm	2D Stacked Mesh NoC	2.6	1.5	4.1
	3D Mesh NoC	2.6	3.0	5.6
	3D Stacked Mesh NoC	2.6	3.0	5.6
	3D Stacked Hexagonal NoC	2.6	1.59	4.19
22 nm	2D Stacked Mesh NoC	1.3	4.95	7.55
	3D Mesh NoC	1.3	9.9	11.2
	3D Stacked Mesh NoC	1.3	9.9	11.2
	3D Stacked Hexagonal NoC	1.3	5.25	6.55
16 nm	2D Stacked Mesh NoC	0.6	8.85	9.45
	3D Mesh NoC	0.6	17.7	18.3
	3D Stacked Mesh NoC	0.6	17.7	18.3
	3D Stacked Hexagonal NoC	0.6	9.38	9.98

La mise en œuvre de style GALS dans cette architecture est basée sur la structure FIFO double horloge représenté dans Figure 14. On utilise une profondeur 4 mots pour le bloc FIFO intégré dans une interface de réseau pour transférer des données du processeur par l'intermédiaire de son maître FLS et d'exploitation du bus esclave à 100 MHz pour le fonctionnement NoC à 333 MHz. pour le processeur de communication NoC, les données provenant du bus du FLS est écrit à la première FIFO double horloge avant d'être mises en paquets pour être envoyée au routeur de transport. En revanche, pour NoC pour le processeur de communication, les paquets arrivent à partir du routeur est d'abord dépaqueté avant d'être écrit dans le FIFO double horloge.

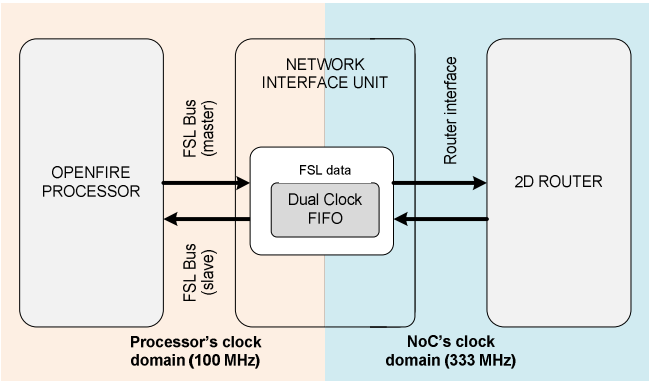


Figure 14: Style de mise en œuvre GALS en utilisant deux horloges FIFO architecture construite dans l'intérieur de la NIU bloc

Pour l'empilement hétérogène, nous avons divisé la conception en 2D dans une dalle de processeur et une autre tuile pour l'architecture NoC comme indiqué dans Figure 15. Le plan du salon et la mise en déroute est montré dans la Figure 16 et Figure 17 en bas et en haut pour les vitesses respectivement. Nous utilisons la technologie Tezzaron de 2 niveau ainsi que le débit de conception similaire expliqué dans le chapitre précédent. Les processeurs sont placés dans la dalle de fond tandis que le CNP est placée dans la dalle supérieure. La liaison verticale est réalisée à partir de signaux d'interface réseau dans le CNP pour le processeur et à la mémoire de données. Par conséquent, tout d'abord nous définissons l'emplacement des microbumps dans l'étage inférieur autour des processeurs et de la mémoire de données, puis nous floorplan le peloton de tête pour l'architecture NoC en plaçant l'interface réseau dans les emplacements microbumps créées à partir de l'étage inférieur d'être aussi proche que possible. Méthodes d'empilement proposée dans [46] n'est pas réaliste parce que le routeur dispose d'une aire relativement petite par rapport au processeur ou tout autres IP core tel que fabriqué dans [47] et [48] où aura une surface de silicon grand vide et ne seront donc modifier le plan d'étage par déplacer le bloc de mémoire d'instructions à l'étage supérieur pour être placé avec l'architecture NoC.

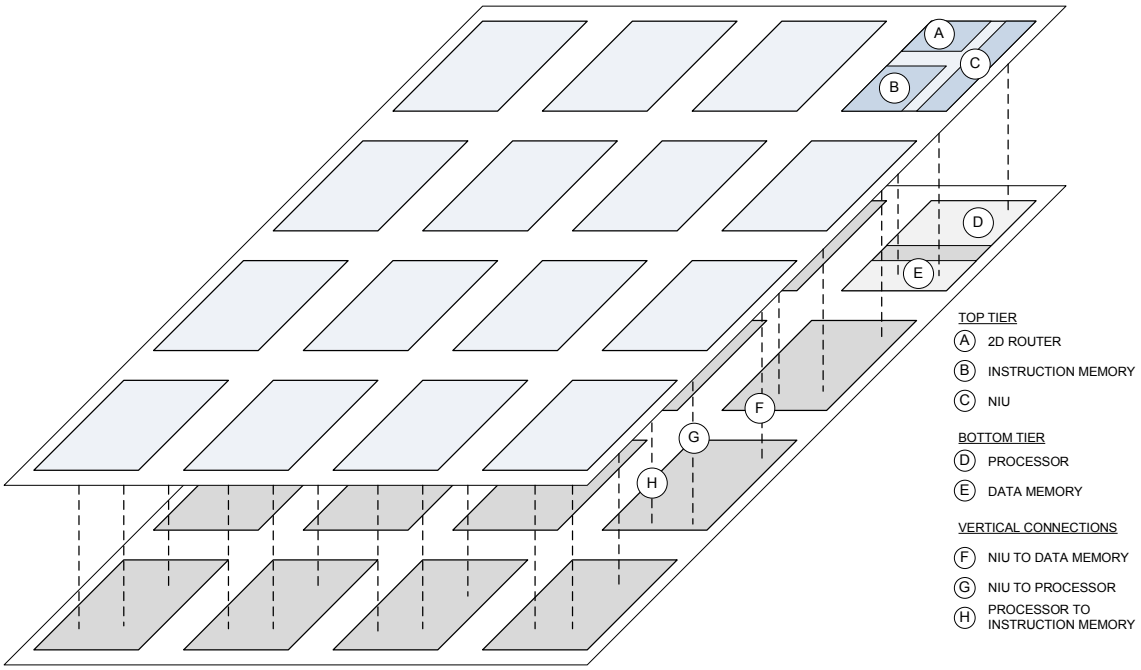


Figure 15: Hétérogène 3D NoC à base de multiprocesseur empilage

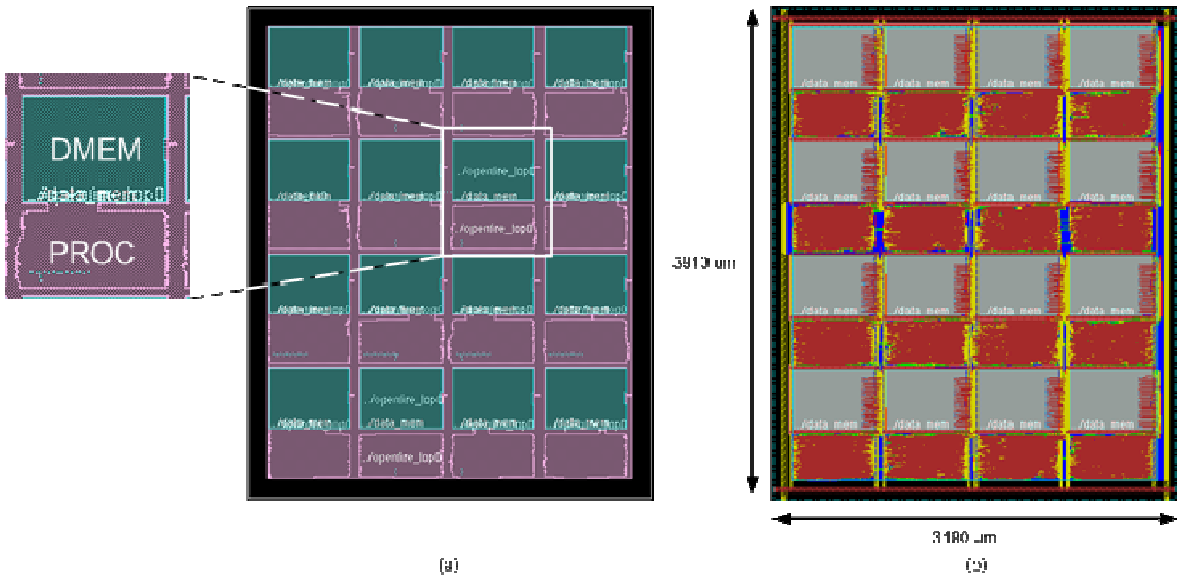


Figure 16: Étage inférieur de la 3D hétérogène d'empilement (a) vue amibes du plan de masse (b) mise en déroute

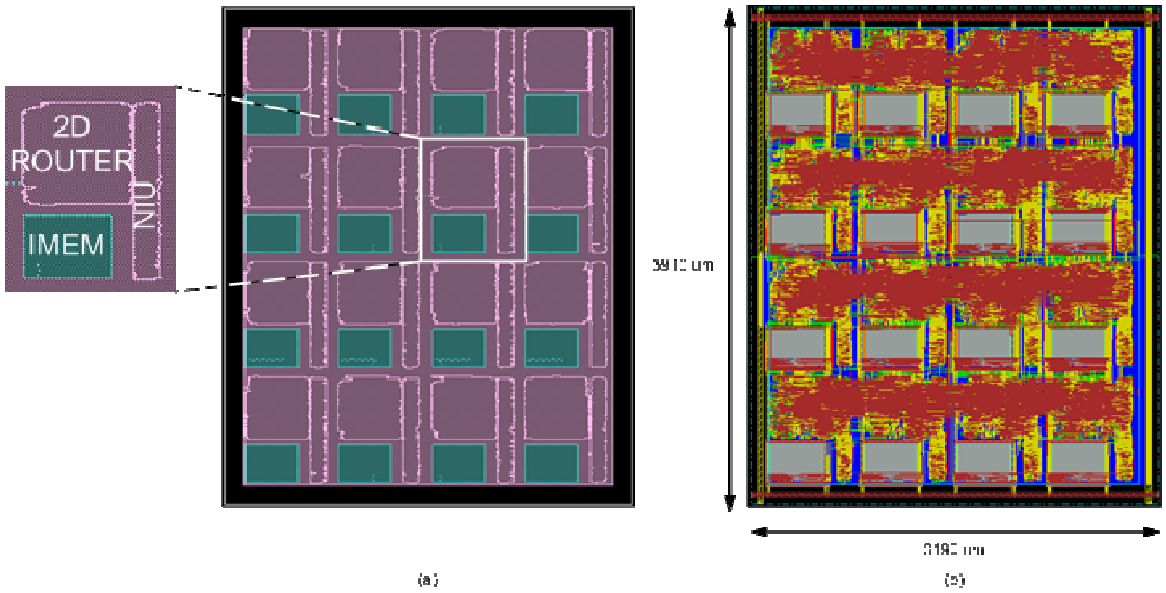


Figure 17: Peloton de tête des hétérogène 3D d'empilage (a) vue amibes du plan de masse (b) en déroute disposition

Il peut être vu de Table 4 qu'il ya presque une réduction de 50% de la superficie de base pour l'empilement 3D hétérogène par rapport à l'architecture 2D en raison de la répartition de l'architecture NoC et de la mémoire d'instructions dans une autre couche. Le nombre de portes est cependant légèrement augmenté au cours de l'architecture 2D principalement en raison des flux optimisation séparée des deux volets au cours de place et pas de route. Sur 188 liaisons verticales par carreau (NIU vers/à partir du processeur et des données mémoire), 70 connexions sont des

connexions FLS processeur alors que le reste est pour les données de connexion et mémoire d'instructions. Nous pouvons également voir la légère augmentation de mesure de longueur totale en 3D hétérogène empilement par rapport à l'architecture 2D en raison de séparer processus d'optimisation 2D lors de place et pas de route.

La comparaison des performances entre la conception 2D et 3D est montré dans Figure 18 où il montre clairement que hétérogène 3D d'empilage améliore légèrement la vitesse de NoC alors aggrave l'horloge du processeur. Une performance accrue de la vitesse NoC est partiellement en raison de la réduction de la zone qui contribue à la réduction de mesure de longueur pour le chemin critique (de l'entrée à enregistrer le chemin). En termes de consommation d'énergie, la consommation a légèrement augmenté architecture de puissance 3D sur l'architecture 2D est due à l'augmentation des portes logiques de l'architecture 3D ainsi que son mesure de longueur totale en raison de la place séparée et itinéraire courir pour chaque niveau.

Table 4: Comparaison des performances de la 3D en 2D et hétérogène empilage

Parameters	2D architecture	3D heterogeneous stacking
Core area (mm^2)	21.4	10.4
Number of gates (million)	2.70	2.73
Number of total *microbumps*	-	3011
Nimber of *microbumps* per tile	-	188
Microbumps for IMEM per tile	-	42
Microbumps for DMEM per tile	-	76
Microbumps for FSL per tile	-	70
Total mesure de longueur (m)	21.1	21.4
Critical path delay for NoC clock (ns)	3.51	3.19
Critical path delay for processor clock (ns)	9.92	10.09
Power Consumption @ 333 MHz (W)	1.38	1.48

Figure 19 montre la répartition horizontale de mesure de longueur 2D MPSoC, bottom tier et top tier du 3D hétérogène où en dessous de 0,8 mm mesure de longueur, on peut constater la réduction du nombre de fil pour l'empilement 3D hétérogène, mais augmentation du nombre de fil lorsque mesure de longueur entre 0,8 mm et 0,9 mm. Comme nous courons endroit séparé et l'itinéraire pour chaque niveau, donc l'outil permet d'optimiser chaque niveau en conséquence, sans considérer

l'architecture 3D complet qui pourrait être l'une des raisons de cette tendance.

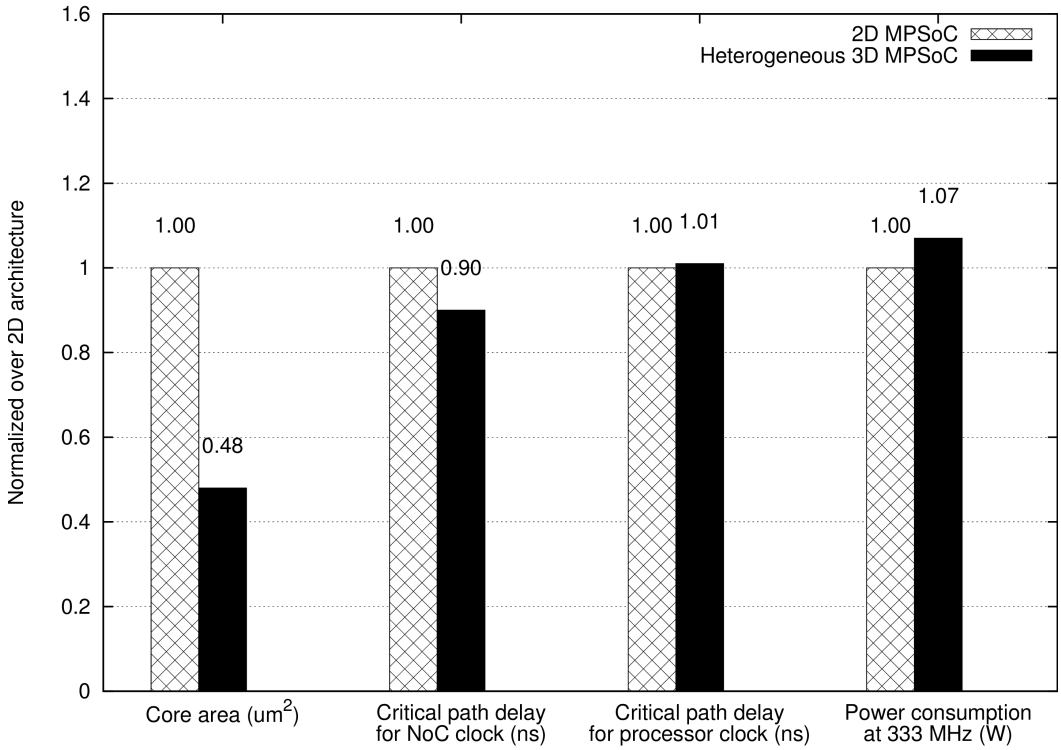


Figure 18: Comparaison des performances 2D et 3D d'architecture MPSoC

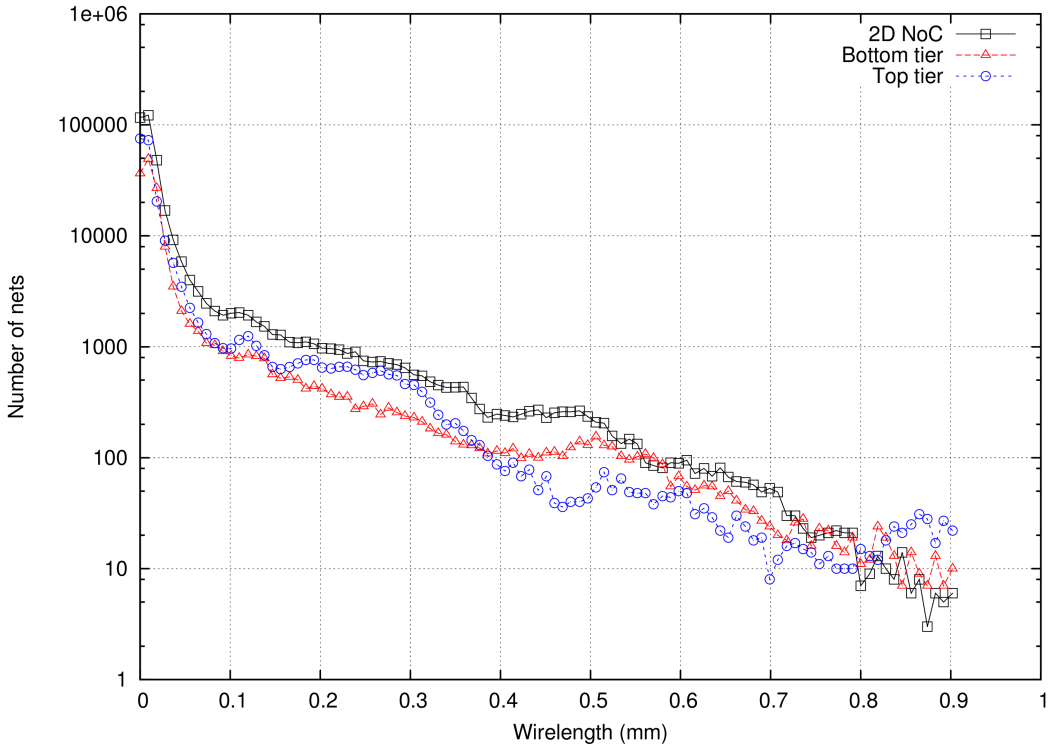


Figure 19: La distribution mesure de longueur horizontale pour MPSoC 2D et 3D MPSoC (bas et haut niveau)

L'horloge de synthèse en arbre pour les deux architectures est faite automatiquement par CTS Engine à SoC Encounter où le fichier de spécification d'horloge est généré en fonction des contraintes temporelles fournies. Un microbumps par signal d'horloge a été placée au centre de l'étage supérieur afin de permettre la distribution équilibre entre les deux niveaux de la source d'horloge qui venant de l'étage supérieur. Comme le montrent les chiffres, CTS Engine synthétisé l'horloge en arbre H-arborescence de la topologie dans les 3 ou 4 premiers niveaux. Table 5 présente la structure de l'horloge de synthèse entre l'arbre de conception 2D et 3D où il est clairement démontré que la structure d'horloge en arbre de conception 3D où se combinent des deux niveaux ont moins de niveau horloge en arbre et le nombre de tampons, même si le nombre de puits est fortement croissante de processeur d'horloge. En revanche, la structure d'horloge en arbre pour la conception de 3D NoC demeure le même nombre de tampons de presque semblable mais de réduire les niveaux de l'horloge en arbre 2 du dessin 2D à cause de l'effet de surface plus petite qui permet de réduire sensiblement la longueur de distribution d'horloge à partir de la source à la sink.

Table 5: Propriétés arborescence d'horloge de structure pour des conceptions 2D et 3D

Parameters	2D		3D (bottom tier)		3D (top tier)	
	Processor clock	NoC clock	Processor clock	NoC clock	Processor clock	NoC clock
Level	17	10	7	-	6	8
Nµmber of buffers	944	1580	879	-	72	1599
Nµmber of sinks	40928	72832	38640	-	2288	72832
Skew (ns)	0.40	0.43	Processor clock skew = 0.76 NoC clock skew = 0.07			

Mesurer la performance de synchronisation de microbumps avec 5 µm et 20 µm pitch emplacement montre l'impact très significatif de microbumps pitch à la performance de synchronisation 3D dans cette conception, même si cette conception a une aire relativement petite pour représenter une conception réaliste. Nous effectuons une analyse temporelle 3D en alimentant les fichiers RC parasites des deux niveaux générés par les SoC Encounter à l'émission L'Heure Synopsys et d'analyser les chemins critiques pour les deux modèles. Table 6 montre la comparaison de lache et décalage de l'horloge entre les deux implémentations où il montre clairement que la conception avec pitch plus grand, mais n'affecte pas le décalage de l'horloge. Même si le synchronisation 3D est également affecté par l'emplacement des microbumps en raison de la longueur de fil horizontale avant d'atteindre les microbumps, de microbumps pitch ne contribue pas significativement aux chemins de synchronisation 3D parce que la différence de taille est relativement petite. Par

conséquent, plus petit microbumps pitch offre une plus grande densité d'interconnexion verticale, mais doivent encore être optimisé par l'architecture cible compte tenu de son emplacement pour les affectations de signaux.

Table 6: Les performances de synchronisation d'emplacements microbumps différentes

Parameters	5 µm pitch	20 µm pitch
Slack	0.08	0.72
Skew	0.52	0.58

Une des principales limitations de l'utilisation des outils de CAO 2D pour la conception et la mise en œuvre de l'architecture 3D IC est le manque de soutien exploration de conception. Pour être en mesure de gagner en performance autant que possible de la technologie 3D, la nécessité d'explorer le design est plus haute importance pour évaluer la différente mise en œuvre compromis pour un matériel ciblé ou une application spécifique avant de procéder à la mise en œuvre de flux de conception complet. Spécifique à la 3D hétérogène d'empilement à niveau des blocs de partitionnement, tant que les chemins critiques résident à l'intérieur de l'architecture bloc en utilisant ainsi des outils de CAO 2D semblent être suffisante pour être en mesure de concevoir ainsi que l'optimisation faisant dû au fait que les outils ne nécessite pas de voir l'architecture 3D complet qui ne possède aucune chemins de synchronisation.

Bien que l'exploration de conception manuelle peut être effectuée en utilisant les outils 2D qui ont été fait pour l'exploration microbumps pitch, ce n'est pas une analyse suffisamment précise parce que généralement, les paramètres de l'architecture 3D est intimement les uns aux autres et doivent donc être fait dans un flux l'intégration 3D complet avec des outils 3D pour une analyse plus précise d'exploration. Par exemple, l'exploration les microbumps pitch pour l'affectation verticale des signaux doit être réalisée simultanément avec l'exploration emplacement microbumps pour être en mesure d'estimer avec précision son impact sur les performances de l'architecture 3D.

References

[1] A. Roy, J. Xu, and M. H. Chowdhury, “Multi-core processors: A new way forward and challenges,” in Microelectronics, 2008. ICM 2008. International Conference on, 2008, pp. 454–457.

[2] S. Borkar, “Design challenges of technology scaling,” Micro, IEEE, vol. 19, no. 4, pp. 23–29, 1999.

[3] S. Kosonocky, T. Burd, K. Kasprak, R. Schultz, and R. Stephany, “Designing in scaled technologies: 32 nm and beyond,” in VLSI Technology (VLSIT), 2012 Symposium on, 2012, pp. 147–148.

[4] T. Thorolfsson, P. D. Franzon, and K. Gonsalves, “Design automation for a 3DIC FFT processor for synthetic aperture radar: A case study,” in Design Automation Conference, 2009. DAC ’09. 46th ACM/IEEE, 2009, pp. 51–56.

[5] B. Vaidyanathan, W.-L. Hung, F. Wang, Y. Xie, V. Narayanan, M. J. Irwin, and M. Gsrc, “Architecting Microprocessor Components in 3D Design Space,” in VLSI Design, 2007. Held jointly with 6th International Conference on Embedded Systems., 20th International Conference on, 2007, pp. 103–108.

[6] Y. Xie, “Processor Architecture Design Using 3D Integration Technology,” 2010 23rd International Conference on VLSI Design, pp. 446–451, Jan. 2010.

[7] R. Weerasekera, M. Grange, D. Pamunuwa, H. Tenhunen, and L.-R. Zheng, “Compact modelling of Through-Silicon Vias (TSVs) in three-dimensional (3-D) integrated circuits,” in 3D System Integration, 2009. 3DIC 2009. IEEE International Conference on, 2009, pp. 1–8.

[8] T. Hsu, K. Chiang, J.-Y. Lai, and Y.-P. Wang, “Electrical characterization of through silicon via (TSV) for high-speed memory application,” in Electronic Manufacturing Technology Symposium (IEMT), 2008 33rd IEEE/CPMT International, 2008, pp. 1–5.

[9] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, “Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs,” Electron Devices, IEEE Transactions on, vol. 57, no. 1, pp. 256–262, 2010.

[10] J. W. Joyner and J. D. Meindl, “Opportunities for reduced power dissipation using three-dimensional integration,” in Interconnect Technology Conference, 2002. Proceedings of the IEEE 2002 International, 2002, pp. 148–150.

[11] J. Ouyang, G. Sun, D. Y. Chen, L. Duan, T. Zhang, Y. Xie, and M. J. Irwin, “Arithmetic unit design using 180nm TSV-based 3D stacking technology,” in 3D System Integration, 2009. 3DIC 2009. IEEE International Conference on, 2009, pp. 1–4.

[12] P. Jacob, A. Zia, O. Erdogan, P. M. Belemjian, J.-W. Kim, M. Chu, R. P. Kraft, J. F. Mcdonald, and K. Bernstein, “Mitigating Memory Wall Effects in High-Clock-Rate and Multicore CMOS 3-D Processor Memory Stacks,” Proceedings of the IEEE, vol. 97, no. 1, pp. 108–122, 2009.

[13] J.-S. Kim, C. S. Oh, H. Lee, D. Lee, H. R. Hwang, S. Hwang, B. Na, J. Moon, J.-G. Kim, H. Park, J.-W. Ryu, K. Park, S. K. Kang, S.-Y. Kim, H. Kim, J.-M. Bang, H. Cho, M. Jang, C. Han, J.-B. LeeLee, J. S. Choi, and Y.-H. Jun, “A 1.2 V 12.8 GB/s 2 Gb Mobile Wide-I/O DRAM With 4 x 128 I/Os Using TSV Based Stacking,” Solid-State Circuits, IEEE Journal of, vol. 27, no. 1, pp. 107–116, 2011.

[14] A. K. Coskun, A. B. Kahng, and T. S. Rosing, “Temperature- and Cost-Aware Design of 3D Multiprocessor Architectures,” in 2009 12th Euromicro Conference on Digital System Design, Architectures, Methods and Tools, 2009, pp. 183–190.

[15] C.-L. Huang, N.-S. Chang, C.-S. Chen, C.-P. Lin, C.-M. Wu, and C.-M. Huang, “A novel design methodology for hybrid process 3D-IC,” in VLSI Design, Automation, and Test (VLSI-DAT), 2012 International Symposium on, 2012, pp. 1–4.

[16] V. K. Jain, S. Bhanja, G. H. Chapman, and L. Doddannagari, “A highly reconfigurable computing array: DSP plane of a 3D heterogeneous SoC,” in SOC Conference, 2005. Proceedings. IEEE International, 2005, pp. 243–246.

[17] S. K. Kim, C. C. Liu, L. Xue, and S. Tiwari, “Crosstalk reduction in mixed-signal 3-D integrated circuits with interdevice layer ground planes,” Electron Devices, IEEE Transactions on, vol. 52, no. 7, pp. 1459–1467, 2005.

[18] D. H. Triyoso, T. B. Dao, T. Kropewnicki, F. Martinez, R. Noble, and M. Hamilton, “Progress and challenges of tungsten-filled through-silicon via,” in IC Design and Technology (ICICDT), 2010 IEEE International Conference on, 2010, pp. 118–121.

[19] M. J. Wolf, T. Dretschkow, B. Wunderle, N. Jurgensen, G. Engelmann, O. Ehrmann, A. Uhlig, B. Michel, and H. Reichl, “High aspect ratio TSV copper filling with different seed layers,” in Electronic Components and Technology Conference, 2008. ECTC 2008. 58th, 2008, pp. 563–570.

[20] G. Katti, A. Mercha, J. Van Olmen, C. Huyghebaert, A. Jourdain, M. Stucchi, M. Rakowski, I. Debusschere, P. Soussan, W. Dehaene, K. De Meyer, Y. Travaly, E. Beyne, S. Biesemans, and B. Swinnen, “3D stacked ICs using Cu TSVs and Die to Wafer Hybrid Collective bonding,” in Electron Devices Meeting (IEDM), 2009 IEEE International, 2009, pp. 1–4.

[21] M. Koyanagi, T. Fukushima, and T. Tanaka, “High-Density Through Silicon Vias for 3-D LSIs,” Proceedings of the IEEE, vol. 97, no. 1, pp. 49–59, 2009.

[22] T. Jiang and S. Luo, “3D Integration-Present and Future,” in Electronics Packaging

Technology Conference, 2008. EPTC 2008. 10th, 2008, pp. 373–378.

[23] K. Nomura, K. Abe, S. Fujita, Y. Kurosawa, and A. Kageshima, “Performance Analysis of 3D-IC for Multi-Core Processors in sub-65nm CMOS technologies,” in Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, 2010, pp. 2876–2879.

[24] R. Yarema, “The Via Revolution,” in 19th International Workshop on Vertex Dectectors - VERTEX 2010, 2010, pp. 1–11.

[25] D. H. Kim, K. Athikulwongse, and S. K. Lim, “A study of through-silicon-via impact on the 3D stacked IC layout,” in Proceedings of the 2009 International Conference on Computer-Aided Design, 2009, pp. 674–680.

[26] E. C. Oh and P. D. Franzon, “Technology impact analysis for 3D TCAM,” in 3D System Integration, 2009. 3DIC 2009. IEEE International Conference on, 2009, pp. 1–5.

[27] P. Gueguen, C. Ventosa, L. Di Cioccio, H. Moriceau, F. Grossi, M. Rivoire, P. Leduc, and L. Clavelier, “Physics of direct bonding: Applications to 3D heterogeneous or monolithic integration,” Microelectronic Engineering, vol. 87, no. 3, pp. 477–484, Mar. 2010.

[28] C.-T. Ko and K.-N. Chen, “Wafer-level bonding/stacking technology for 3D integration,” Microelectronics Reliability, vol. 50, no. 4, pp. 481–488, Apr. 2010.

[29] L. Di Cioccio, I. Radu, P. Gueguen, and M. Sadaka, “Direct bonding for wafer level 3D integration,” in IC Design and Technology (ICICDT), 2010 IEEE International Conference on, 2010, pp. 110–113.

[30] J. J. McMahon, E. Chan, S. H. Lee, R. J. Gutmann, and J.-Q. Lu, “Bonding interfaces in wafer-level metal/adhesive bonded 3D integration,” in Electronic Components and Technology Conference, 2008. ECTC 2008. 58th, 2008, pp. 871–878.

[31] R. S. Patti, “Three-Dimensional Integrated Circuits and the Future of System-on-Chip Designs,” Proceedings of the IEEE, vol. 94, no. 6, pp. 1214–1224, 2006.

[32] B. Black, M. Annavaram, N. Brekelbaum, J. Devale, L. Jiang, G. H. Loh, D. McCauley, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb, “Die Stacking (3D) Microarchitecture,” in Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on, 2006, pp. 469–479.

[33] R. Weerasekera, L. Zheng, D. Pamunuwa, and H. Tenhunen, “Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs,” in Computer-Aided Design, 2007. ICCAD 2007. IEEE/ACM International Conference on, 2007, pp. 212–219.

[34] M. Santarini, “Thermal integrity: a must for low-power-IC digital design,” EDN, 2005. [Online]. Available: www.edn.com.

[35] K. Puttaswamy and G. H. Loh, “Thermal Herding: Microarchitecture Techniques for

Controlling Hotspots in High-Performance 3D-Integrated Processors," in High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on, 2007, pp. 193–204.

[36] E. Wong and S. K. Lim, "3D Floorplanning with Thermal Vias," in Design, Automation and Test in Europe, 2006. DATE '06. Proceedings, 2006, vol. 1, pp. 1–6.

[37] J. Yang, K. Athikulwongse, Y. Lee, S. K. Lim, and D. Z. Pan, "TSV stress aware timing analysis with applications to 3D-IC layout optimization," in Design Automation Conference (DAC), 2010 47th ACM/IEEE, 2010, pp. 803–806.

[38] M. C. Hsieh, Y.-Y. Hsu, and C.-L. Chang, "Thermal Stress Analysis of Cu/Low-k Interconnects in 3D-IC Structures," in Microsystems, Packaging, Assembly Conference Taiwan, 2006. IMPACT 2006. International, 2006, pp. 1–4.

[39] P. D. Franzon, W. R. Davis, and T. Thorolffson, "Creating 3D specific systems: Architecture, design and CAD," in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010, 2010, pp. 1684–1688.

[40] D. V Campbell, "Yield modeling of 3D integrated wafer scale assemblies," in Electronic Components and Technology Conference (ECTC), 2010 Proceedings 60th, 2010, pp. 1935–1938.

[41] G. Smith, L. Smith, S. Hosali, and S. Arkalgud, "Yield considerations in the choice of 3D technology," in Semiconductor Manufacturing, 2007. ISSM 2007. International Symposium on, 2007, pp. 1–3.

[42] E. J. Marinissen, "Testing TSV-based three-dimensional stacked ICs," in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010, 2010, pp. 1689–1694.

[43] T. Thorolfsson, G. Luo, J. Cong, and P. D. Franzon, "Logic-on-logic 3D integration and placement," in 3D Systems Integration Conference (3DIC), 2010 IEEE International, 2010, pp. 1–4.

[44] T. Thorolfsson, N. Moezzi-Madani, and P. D. Franzon, "Reconfigurable five-layer three-dimensional integrated memory-on-logic synthetic aperture radar processor," Computers & Digital Techniques, IET, vol. 5, no. 3, pp. 198–204, 2011.

[45] M. Pathak, Y.-J. Lee, T. Moon, and S. K. Lim, "Through-silicon-via management during 3D physical design: When to add and how many?," in Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on, 2010, pp. 387–394.

[46] V. De Paulo and C. Ababei, "3D Network-on-Chip Architectures Using Homogeneous Meshes and Heterogeneous Floorplans," International Journal of Reconfigurable Computing, vol. 2010, pp. 1–12, 2010.

[47] S. R. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, “An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS,” Solid-State Circuits, IEEE Journal of, vol. 43, no. 1, pp. 29–41, 2008.

[48] M. Grange, A. Y. Weldezion, D. Pamunuwa, R. Weerasekera, Z. Lu, A. Jantsch, and D. Shippen, “Physical mapping and performance study of a multi-clock 3-Dimensional Network-on-Chip mesh,” in 3D System Integration, 2009. 3DIC 2009. IEEE International Conference on, 2009, pp. 1–7.

CHAPTER 1
INTRODUCTION

Electronic designs have been growing rapidly for the past several years triggered by the rising demand of smartphones and tablets in the market. The high-end mobile devices increase the demand for more functionality as well as computing power to be able to run many more applications. Small form factor, higher performance and lower power are among the important requirements for the mobile devices in order to deliver smaller, cheaper and faster consumer electronic devices. Figure 1 shows the trend for the number of processing elements in SoC consumer portable devices according to the International Technology Roadmap Semiconductor (ITRS) [1]. As shown in the figure, in the near future, the number of processing elements is expected to increase to more than 100 processors. Additionally, the memory size is also projected to increase dramatically in the future along with the increasing number of processing elements.

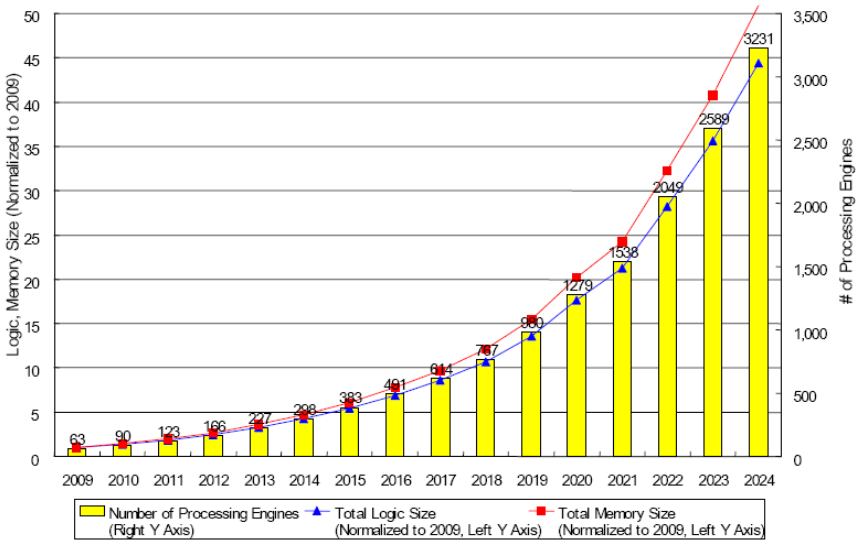


Figure 1: Number of processing engine and logic memory size trends in electronic systems

Increasing the performance of single processor design through higher clock frequency poses drawback of higher power consumption and thus multiprocessor architecture has been introduced where a design has several simple processors that run at a lower frequency and lower voltage supply. Power can be reduced by shutting down idle processors depending on which applications are running, while the performance can be improved through parallel execution of applications using multiple processors. We are moving from computation-extensive architecture to communication-extensive architecture introduced from this multiprocessor architecture. However, designing high performance multicore processor architecture requires several challenges to be

solved such as interfacing different IP cores, design automation, verification and software programming [2]. In order to meet the demand of communication requirement, Network on Chip (NoC) is developed to overcome the limitations of bus based architecture [3]. Some of the disadvantages of bus architecture are long signal delay due to arbitration policy and complex wiring that contributes to the overall power consumption of the system. In contrast, NoC offers scalability and huge amount of communication bandwidth when increasing the number of processors to perform complex operations [4]. In addition, it supports parallel communication between different processing elements and also improves communication speed as it does not require global arbitration policy.

Initially, we rely on the CMOS scaling features to get more performance which is achieved by reducing the physical dimension of the transistor so that many more transistors can be packed into a single chip, thus increasing the performance through more deeply pipeline architecture. CMOS transistor scaling, is a technique to increase the performance of $1/k$ at constant power density by reducing oxide thickness (t_{ox}), transistor gate length (l), and transistor gate width (w), where k is the scaling factor. This is true until we reach 130 nm technology because for the following smaller technology node, enhancers have been added during fabrication processing steps to make sure that the transistor can be operated at the desired performance. However, moving towards smaller process technology introduces many great economical and technological challenges and at the same time decreasing performance benefits at every scaling nodes [5]. For example, in 90 nm and 65 nm technology, strain has been added, while for 45 nm and 32 nm technology, strain, low-k dielectric for inter-layer metal insulation and high-k dielectric metal gate have been used to control transistor's integrity [6] and many more enhancer methods will be needed as we move towards sub-20 nm technology. On the other hand, the limitation of CMOS scaling such as maximum voltage limits and device variability have also impacted design techniques at system and circuit level where additional design techniques are required to enable increasing performance improvement and power reduction with cost reduction [7].

Consequently, 3D IC technology enables higher device integration and improves design performance by stacking wafers or dies on top of the other and interconnected using TSV technology [8]. It has been studied by several researchers for the past few decades but only now has gained great attention where it is seen as a promising solution now as many people realize that 2D scaling is becoming more and more difficult to manufacture. Using this technology, the thin wafers or dies are stacked in several layers as depicted in Figure 2 as an example and then is packaged using conventional packaging methodology. This new technology offers potential benefits of faster

speed, lower power consumption, integration of heterogeneous technology, smaller form factor and high device integration density. In contrast to CMOS technology scaling, 3D integration is a promising solution to drive the future of VLSI circuits to support the continuous demand of high performance electronic systems. These benefits will be described in details in the later sections.

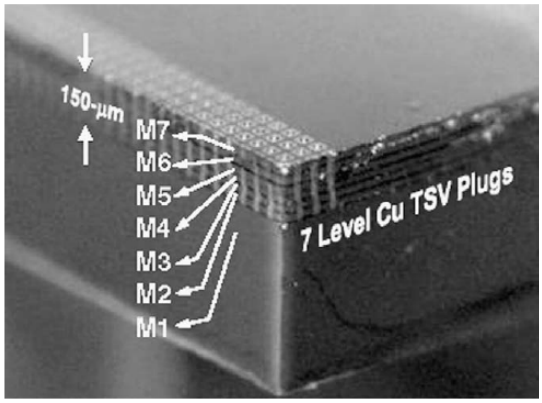


Figure 2: 3D integration example showing seven stacks of wafer connecting using Cu TSV [6]

In 3D integration, the long interconnect wire length is reduced to square root of its original length in 2D architecture due to the short vertical connection using either TSVs or microbumps. This improves the speed where it reduces the RC delay of the long interconnect wire in 2D architecture and also reduce the number of buffers along the interconnect wire wherein eventually power consumption is decreased as well. 3D integration also supports integration of heterogeneous technology such as digital, analog, RF and MEMS technology where they can be fabricated according to their optimal process technology and then stacked with other process technology. Finally 3D integration introduces design miniaturization enabling higher density memory or logic capacity.

Despite the benefits brought by this technology, it also faces several critical obstacles such as thermal issues, power delivery structure and clock tree synthesis. As 3D integration stacks several active silicon layers vertically, the device density per unit volume is increased and thereby increases the power density. Together with the long thermal transfer path between multiple stacked dies and poor thermal conductivity of dielectric layer in the multiple stacked dies cause the temperature in the chip to arise. Another effect is that there exist thermal hotspots due to the different power consumption by different logic blocks in each layer. This generates thermal gradient over the chip which create variations that could possibly affect the integrity and reliability of the devices over time. Power delivery structure also needs to be revised in order to ensure enough current supply to the farthest tiers in the stack and clock tree network must be optimized to gain maximum

performance out of 3D stacking where TSV plays a crucial role for electrical connections between tiers. Although some of the testing techniques for 2D architecture can be extended to enable testing of 3D architecture such as built-in self test (BIST) and boundary scan method, however the manufacturing of TSV introduces new defect mechanism such as shorts or opens due to misalignment and micro-viods and therefore requires a new approach to test these defects. These are some of the main hurdles in 3D integration technology that need to be solved before it can be commercially viable in many products.

1.1 Research Motivations

As this is a relatively new technology at the moment, many issues are still being researched extensively in industry and academics. Looking at the publication in conferences and journals, we realized that there is a lack of real design and implementation being carried out to have more realistic results on the performance analysis of this technology. In pacticular, research works in 3D NoC architecture have been done mostly using software simulation using cycle-accurate simulator which provide unrealistic results and thus are not sufficient to evaluate the pros and cons of 3D technology. Therefore, one of the main motivations in this work is to conduct performance analysis by using real design and implementation of 3D technology using technology that are available to use. Through this approach, we aim to have more realistic results and thus could help us to better understand the trade-off offers using this technology.

A part from that, we focus our work on the architectural aspect of this technology as we have specifically chosen to use Tezzaron two-tier 3D technology for the technological aspect. Current electronic devices mostly have more than one processing core in order to have more capability to run various applications with higher performance. This is due to the fact that the performance of a single processor architecture cannot be improved anymore because of the power wall and the memory wall. Therefore, it is interesting to know how this 3D technology could be used to overcome the multiprocessor issues that are being faced now to be able to improve its performance. Several works have been done performing design and implementation on multiprocessor architecture using 3D technology where a handful number of works showed quite significant performance improvement using 3D technology when compared to 2D architecture. However, none of the previous works analyzed the 3D NoC architecture performance base on real design and implementation results, which is the main objective of this work. It is interesting to understand what kind of NoC topologies (2D or 3D topologies) is better in 3D architecture in terms of performance where we do not have to consider it when designing a 2D architecture. On the other hand, research

works on 3D GALS-based multiprocessor architecture have also not been reported especially concerning physical design implementation issues. This is the reason that motivates us to design and implement this type of architecture to be able to conduct performance analysis when compared to its 2D implementation and identify physical design issues that need to be taken into consideration during the design process.

A common approach to design 3D architectures is to use the state of the art 2D EDA tools as has been reported by many works in the literature since there are no commercial true 3D design tools available in the market to date. This approach does not guarantee maximum performance gain of the 3D architecture over its 2D architecture because 2D EDA tools are not able to see the complete 3D architecture to be able to perform necessary optimization processes to achieve the target performance requirements. The lack of study in the literature regarding what is the impact of 2D EDA tools on the 3D architecture performance motivates us to conduct a design space exploration experiment of 3D MPSoC architecture to investigate this issue in details.

1.2 Summary of Arguments

1.2.1 Deep Understanding about the Target Architecture to Maximize Performance Improvement

In this thesis, although many previous published papers have shown that 3D IC technology gives many advantages in terms of area, performance and power when implemented using 2D design tools, we argue that there is still a need for the careful architectural decisions in order not to diminish the potential benefits it could offer. Bad architectural implementation choices such as bad partitioning methods (with manual and not optimize partitioning), bad TSVs/microbumps planning and bad 3D specific architecture (such as 3D NoC topology vs 2D NoC topology) particularly for heterogeneous 3D stacking can only make the performance worse than its 2D architecture or if not worse, will only give marginal improvement to be able to justify the devoted design efforts and time for designing in 3D technology. Deep understanding about the target architecture to be implemented as well as the target 3D IC technology to be used are completely essential to help making the right decision on choosing various 3D specific parameters such as power delivery methods, clock tree structure, TSVs count and location, power and thermal management methods and design-for-test consideration.

1.2.2 Process Technologies for 3D IC Technology Depending on the Different Target Implementations

On the other hand, although the current 3D IC technologies that are available today such as Tezzaron two-tier technology using bulk CMOS 130 nm technology of Global Foundries, IMEC technology using bulk CMOS 130 nm technology and MIT LL three-tier technology using FDSOI 180 nm technology, we argue that implementing 3D IC technology using advanced process technology such as 45 nm is more appealing for high performance computing due to the fact that in advanced technology global delay is heavily dominated by wire delay propagating through high number of metal interconnect layers thereby will have stronger impact on the performance improvement of the 2D architecture than in older process technology. Using heterogeneous 3D IC technology that stacked different architectures that have been optimized in their own matured process technology is a more attractive solution to get higher performance improvement. However, using old process technology for 3D architecture implementation can be considered for low-end devices to improve performance of previously developed 2D devices due to the cost issues when migrating to the other process technologies.

1.2.3 3D-aware EDA Tools with 3D Optimization Capability for Designing 3D IC Technology

From the design perspective, true 3D aware tools for designing 3D architectures that are capable of doing 3D aware design and optimization such as 3D synthesis, placement, clock tree synthesis and routing under various design constraints such as timing, power and thermal constraints are absolutely needed to be able to gain the highest possible performance improvement from 3D technology. Although current matured 2D EDA tools can be used to design 3D architectures with some design flow modifications customized to the specific 3D IC technology target, we argue that the performance improvement is still limited to the architectural implementation choices as mentioned earlier and we do not have full control over architectural parameters of a complete 3D design to do an architectural analysis before making decisions on the best architecture to be implemented whereby the 2D tools are not aware of the complete 3D design when optimizing it to meet the design constraints especially tight design constraints for high performance arhitectures. Furthermore, designing 3D architecture using 2D tools requires much more efforts whereas if using true 3D-aware tools, it will be more productive as it allows designers to concentrate more on the 3D architectural issues for a given particular target device rather than spending time and resources configuring the 2D tools or developing design flows for every 3D architecture projects.

1.3 Thesis Contributions

The contributions of this thesis are summarized as follows:

1. Proposed novel design flow for 3D architecture design using available 2D EDA tools primarily focusing on 3D timing verification. This 3D timing verification is possible because vertical inter-tier connection is created using microbumps which has negligible delay in this particular 3D technology. The proposed 3D design method has been used for the experiments in this thesis conducted to study various architectural implementations that are feasible using 3D technology. As will be explained in detail in Chapter 2, the proposed methodology offer several advantages such as early and more accurate 3D performance estimation through post synthesis netlist timing analysis and allow faster 3D architectural exploration to find an optimized architecture to obtain the most benefit from stacking dies.
2. Presented NoC architectures exploration in 3D architecture through physical design implementation motivated from the earlier studies in the literature that carried out performance analysis from software implementation. We designed and implemented 3D Mesh NoC architecture and 3D Stacked Mesh NoC architecture using homogenous stacking of 2D router architecture in two tiers using Tezzaron 3D technology and compared its performance with the 2D NoC architecture. We proposed a new topology for 3D NoC architecture which is hexagonal topology that provides better performance than the other architectures in 3D technology implementation due to the equal inter-router wire length.
3. Conducted heterogeneous 3D stacking implementation for GALS multiprocessor architecture by stacking NoC architecture on top of the processor due to the limited number of works in 3D architecture based on GALS implementation. Among the benefits of this stacking method are better control of thermal and power management methods due to the separate physical layer of the architecture having different thermal and power consumption profile. In this study, we analyzed the performance of heterogeneous 3D stacked architecture that have been manually partitioned into two tiers and compared with its correspondent 2D architecture to identify design trade-off as well as physical design implementation issues. We also conducted analysis on wire delay effect to the 3D NoC architectures performance by doing physical design implementation through two standard cell libraries representing old and advanced process technologies.
4. Carried out a design space exploration of 3D MPSoC architecture to analyze the impact of 2D EDA tools to its performance such as timing quality, power consumption and wirelength. Since it is understandable the limitation of using 2D EDA tools to design and implement 3D architecture, this study investigated the performance impact of 3D MPSoC architecture when

2D EDA tool options, in particular the placement and routing options is varied enabling us to understand several important implementation issues that have not been pointed out previously. We focus on timing and power optimization options in the 2D EDA tools for the exploration because both metrics are among the most essential parameters that are considered during the 3D architecture design.

1.4 Thesis Organization

This thesis is organized as follows: In the next chapter (Chapter 2), we present our initial work on 2D NoC-based MPSoC design and implementation to identify design issues related to the MPSoC architecture before embarking on the 3D IC research work.

Chapter 3 briefly explains the overview of 3D IC technology by firstly introducing the issues being faced by the current 2D architecture. Then we look into different aspects of 3D technology including TSV structure, bonding methods, stacking orientation and several other challenges that need to be overcome before this technology can be adopted as a mainstream technology. The 3D standards are also presented highlighting the need for standards in various aspects of 3D technology because through standardization it would make adoption of this technology faster. Finally we present the state of the art 3D architecture implementations that have been completed to date targeting various design objectives revealing the concrete proof of the benefits of this technology that have been discussed earlier in the chapter.

Chapter 4 discusses a proposed design methodology specific to the Tezzaron 3D technology two-tier face-to-face integration but architecturally generic. This design flow takes benefits from the small structure of microbumps having small/negligible delay for the vertical inter-tier connections. Therefore, performing 3D timing analysis at post-synthesis stage allows us to verify the 3D design early in the design stage rather than post-place and route stage. After analyzing the timing path in the 3D design, necessary modification in the RTL can be done to achieve its target performance without having to wait until finish place and route the design that save a certain amount of time.

Chapter 5 explains about the exploration of 3D NoC architecture through design and implementation using Tezzaron two-tier technology. Performance analysis is conducted based on the routed netlists. Current 2D EDA tools have been used for the implementation based on the methodology explained in Chapter 4. We compare the 3D NoC architecture implementations in order to obtain the best topology to be used for the 3D technology. Two standard cells technologies

have been used for the implementation to have better understanding about the effect of wire delay to the performance of 3D architectures. We also present our target 3D IC designs being developed to be sent for fabrication comparing two MPSoC designs developed by two teams, the GIPSA-Lab team and the ENSTA ParisTech team. The MPSoC architectures used different NoC topologies in order to measure their performance in real implementation when running applications and also to study various 3D implementation issues.

Chapter 6 describes the design and implementation of heterogeneous 3D stacking MPSoC architecture employing GALS style to analyze various architectural trade-off. This chapter features another feasible 3D NoC-based MPSoC implementation through heterogeneous implementation by stacking different architectures on different layers using similar process technology. We conduct several analyses regarding the clock tree structure and critical paths between 2D MPSoC and heterogeneous 3D MPSoC to highlight design and implementation issues with respect to the use of 2D EDA tools. Additionally, we perform experiment on varying microbumps pitch and location for the vertical connection and study its implication on the performance of the 3D architecture identifying the limitations it could impose for designing complex 3D application when using 2D EDA tools.

Chapter 7 presents a design space exploration of 3D MPSoC architecture using 2D EDA tool to analyze the impact of EDA tool on 3D architecture performance as well as to highlight design issues related to designing 3D architecture. Since true 3D design tools are not available until now, specifically design tools that are capable of doing 3D physical design as well as optimization, this exploration allows us to have better understanding about how different options in the EDA tools lead to different optimization results of 3D architecture.

Chapter 8 concludes the works presented in this thesis. It also highlights main contributions reported in this thesis to obtain more understanding of 3D integration focusing on architectural point of view. Future works based upon the works conducted in this thesis are also proposed for futher investigation of 3D architecture design with complete analysis.

CHAPTER 2
2D NOC-BASED MPSOC DESIGN AND IMPLEMENTATION ON FPGA

The era of multiprocessor system-on-chip (MPSoC) has brought a new challenge for modern electronic systems. Communication between IP cores and other peripheral in the MPSoC environment is becoming critical which will affect the performance. Network-on-Chip (NoC) is a promising solution for MPSoC communication limitation. Several NoC studies have been reported over the years but only a few discussed about the actual hardware implementation. In this chapter, we presented FPGA design and implementation of MPSoC system with NoC architectures in order to obtain its actual performance. To improve design productivity, we use Arteris design tool to automatically generate NoC architectures and also supports various interface protocols to other IPs. A case study of Discrete Cosine Transform (DCT) using parallel programming is carried out to validate the design. The goal of this chapter is to present our initial work regarding the 2D NoC-based MPSoC implementation on FPGA which enable us to identify design issues for NoC implementation before proceeding with the 3D IC design process.

2.1 Introduction

MPSoC has been emerged over many years in response to the need of embedded computing requirements such as handheld devices and laptop computers. This is in contrast to the multicore processors that do not have tight requirement commonly use in desktop computer and server applications. The need for programmability, high performance in real time application and low power operation are among the main motivation of MPSoC for various embedded system applications which includes multimedia, signal processing and automotive [9]. As embedded systems demand high performance to support multiple functions, large scale MPSoC design has been emerging. With the current IC technology allows us to use million of transistors in a chip, large scale MPSoC is possible. However, the performance of large scale MPSoC may be degraded due to the communication efficiency between processors and other IPs.

Traditional bus based architecture and dedicated interconnection managing communication in the MPSoC system face several drawbacks such as less scalable, complex wiring connection which contribute to large power consumption, low performance due to arbitration scheme and less design space exploration. Shared bus interconnection has limitation in its scalability because all bus accesses must be serialized by the arbitrator. Bus structure cannot handle in environment such as large number of request bus and higher bandwidth interconnection. Bus structure has also limitation

on large wiring delay when large number of components attached to the bus due to the physical capacitance of the bus wires grows. Shared bus offers system reusability and available bandwidth is shared among nodes but reduce operating frequency with system growth. Advantage of bus architecture is enhanced communication performance but poor reusability where dedicated channel cannot scale well with system complexity [10]. Advance bus architectures are also were proposed such as ARM AMBA [11], OpenCore WISHBONE system on chip interconnection [12], ST Microelectronic Bus (STBus) [13] and IBM CoreConnect [14] as the extended version to achieved high performance of bus architecture. Advanced bus architecture adopt hierarchical structure in order to get scalable communication throughput and partition communication domains into several group of communication layers depending on the bandwidth requirement such as performance.

Networks-on-chip (NoC) provides solution to the limitation of bus based architecture and dedicated interconnection scheme in multiprocessor system-on-chip (MPSoC) design. NoC architecture offers scalability and flexibility of MPSoC design to achieve better performance as well as supporting large scale MPSoC design. By facilitating NoC in the MPSoC system, adding additional elements in the system is not requiring too much effort. It is also can be used to integrate different type of components as the NoC architecture is only dependant on the protocol for its interface. Due to less complex wiring, NoC improve the MPSoC design by using less hardware area, better performance, and also less power consumption because of shorter wiring distance between components [15]. Network on chip provide scalability and freedom from the limitation of complex wiring. Using NoC, wiring for the interconnection is shorten. NoC reduce SoC manufacturing cost, SoC time to market, SoC time to volume, increase SoC performance. NoC also increase system throughput. NoC offers high flexibility and regularity of a network structure supporting simpler interconnect models and greater fault tolerance. NoC able to integrate many different IP cores such as processors, DSP cores, memory blocks, FPGA blocks, dedicated hardware. NoC provide good solutions in numerous applications [16] such as flexible product that should be reconfigurable and programmable, applications with heterogeneous task mix, design which are basis of several product variants, applications with stringent time-to-market requirements, products where reuse at the block, function and feature level is considered valuable.

In this work, we demonstrated MPSoC design with 16 MicroBlazes as masters and 16 BRAMs as slaves. The masters and slaves are connected through 2-ary 4-tree NoC architecture. Several interfaces have been designed to accommodate different communication standard between MicroBlaze and BRAM with the NoC. Additional Application Programming Interface (APIs) is also developed used for synchronization of the masters. For the evaluation methodology, parallel

programming for Discrete Cosine Transform (DCT) application is tested on the design based on several image size and different MicroBlaze configuration.

2.2 Related Works

Mesh and Torus topology have been the popular choice for FPGA implementation because of it's simple routing algorithm and easy implementation on hardware. Various configuration of master and slave combination has also been considered in previous work. In [17], FPGA implementation of Torus topology is presented and proposed new router architecture and algorithm to solve congestion problem from Mesh topology. Design and implementation on Altera Stratix II FPGA for 2D Mesh architecture is presented by [18] with the aim to evaluate scalability of Mesh network by experimenting different number of processing elements. Another 2D Mesh architecture prototyping in FPGA is reported in [19] where they evaluate Mesh NoC to compare with shared-bus and point-to-point architecture. In [20], 3 x 3 Mesh network was designed and implemented and has been tested for Charge-Coupled Device (CCD) application. They designed custom router architecture based on circuit switch protocol. Other than regular NoC topologies, work on custom NoC topologies have also been developed such as SUNFLOOR [21] and SPIN [22]. Work on parallel implementation for DCT has also been presented in the literature. In [23], MPSoC design using custom system level design framework called Deadalus was presented by implementing JPEG-based image compression to Xilinx FPGA. Different number of MicroBlaze processors as well as dedicated DCT cores, which is up to 16 MicroBlazes and 8 DCT IPs were designed with exploiting task and data parallelism for the target JPEG application. The speed up of almost 20 times is achieved using combination of MicroBlaze processors and DCT cores. Homogeneous MPSoC system for JPEG application also has been reported in [24] implemented on Xilinx V4 LX25 FPGA. The MPSoC comprises up to four MicroBlazes which achieved the speed up of three times. Another work proposed parallel implementation of DCT on two DSP processors which shown increasing speed up calculation time [25]. This work proposed MPSoC based on 16 MicroBlazes implemented on Xilinx V4 LX200 FPGA. Several EDA tools have been used throughout the design to improve design productivity for designing large and complex system. Furthermore, NoC architecture is used for communication within the MPSoC system.

2.3 EDA Tools Integration

EDA tools play an important role in the development of electronic design to achieve various objectives such as meeting performance requirements and reducing time-to-market. In this work, we

use several EDA tools to complete the design and execution of FPGA which are Arteris NoCcompiler version 1.12 from Arteris [26], Xilinx ISE and EDK 9.1 from Xilinx [27], and ZeBu Compiler from EVE [28].

NoCcompiler is a NoC configuration environment for Arteris NoC IP Library. Arteris NoC Transport and Transaction Protocol (NTTP) is packet-based NoC architecture. In order to provide communication between IP blocks over a NoC, the NTTP uses three layer approaches which are transaction, transport and physical layers. Transaction layer has Network Interface Units (NIUs) such as AMBA High Performance Bus (AHB), AMBA Advanced Extensible Interface (AXI) and Open Core Protocol (OCP) that define exchange of information between NIU to perform certain transaction. For OCP interface, the NIUs are compliant with OCP 2.2. Each master (or initiator) and slave (or target) are connected to a NoC using OCP-to-NTTP and NTTP-TO-OCP NIUs respectively through a socket. OCP-to-NTTP NIUs allow master to be connected to the NoC by translating OCP transactions into equivalent NTTP packet sequence and vice versa NTTP-to-OCP NIUs. It supports OCP data bus of 32, 64 and 128 bits which is manually specifies by designer. Transaction between NIU and IP blocks can be of request or response. Most transactions are in two steps; a master sends request packets and a slave response packets. In the transport layer, the packets are routed through the NoC using Packet Transport Units (PTUs) such as switches, adapters, converters and others. Request packets can be 33 or 36 bits for data cell while for response packets the data cell size is always 33 bits.

Next, physical connection of the packet in the NoC is defined in the physical layer. Switch is an essential element of the NoC. It receives packets from input ports and forwards each packet to a specific output port. In this work, the switch is based on the Arteris Danube IP Library. It uses wormhole routing algorithm to reduce latency and has full throughout arbitration due to one routing decision per input per cycle. NoC architecture from NoCcompiler can be exported to synthesizable RTL files either in VHDL, SystemC or Verilog. The same flow has been used previously in different works [29]. Xilinx Embedded Development Kits (EDK) is a set of tools and Intellectual Property (IP) for developing embedded processor system targeting Xilinx FPGA devices. EDK provide an environment for designing complex embedded systems combining hardware blocks and software applications. Processors and other peripherals are connected using On-Chip Peripheral (OPB) or Processor Local Bus (PLB). Within the EDK environment, as there is no OCP protocol support by this tool, two IPs have been developed to interface NoC architecture with Microblaze processor and BRAM. For the MicroBlaze processor, it used FSL interface because it is simple to modify in order to design interface to the OCP protocol. The EDK tool is used to build complete

MPSoC system with NoC architecture that can have different number of masters (processors) as well as slaves (memories).

Open Core Protocol (OCP) has been used to interface between MicroBlaze (masters) and BRAM (slaves) to the NoC. It is defined by an international committee, OCP-IP [30]. The OCP interface signals for master and slave are shown in Table 1. It provides independence from bus protocols and allows us to develop reusable IP cores without having loss of high performance access to the NoC. The OCP protocol uses 32 bit data width for masters as well as slave connection. The generated VHDL file of the NoC from NoCcompiler is simulated using ModelSim by verifying these signals, in order to validate the NoC RTL file before it is integrated in the EDK environment. The command *MCmd* defines the operation to test. This signal is important to be used in the synchronization of masters for parallel programming application. Masters begin by writing data stored in the field *MData* in the specific slave. We can verify if this operation was done successfully by testing *MDataValid* signal. A master can also read data from a slave represented by the signal *SData*.

2.3.1 Design Flow

The design flow of this project is shown in Figure 3. The NoC architectures are first designed using NoCcompiler. Internal NTTP proprietary protocol and interface units protocols are configured then the architecture is realized using switches and route tables. Once the design is finished, it is tested to verify the connectivity and RTL file in VHDL is generated describing the NoC. The generated VHDL file is then integrated in the EDK as an Intellectual Property (IP) peripheral with the MPSoC system consist of processors and other peripherals such as memory. A simulation is also carried out using ModelSim to verify the design. With the NoC VHDL code imported as an IP to the EDK environment, the MPSoC design can take place and software applications are also developed in this phase. The NoC connection (with its OCP interface unknown to EDK) is realized through direct access to MHS hardware description file. At the end of this phase two types of files are generated: Netlist (.ngc) files for hardware description and executable (ELF) files for MicroBlaze processors.

Table 1: OCP-IP interface signals

Signals	Driver	Width	Function
Clk	varies	1 bit	clock input
EnableClk	varies	1 bit	enable OCP clock
Maddr	master	configurable	transfer address
MCmd	master	3 bits	transfer command
Mdata	master	configurable	write data
MdataValid	master	1 bit	write data valid
MrespAccept	master	1 bit	master accepts response
ScmdAccept	slave	1 bit	slave accepts transfer
Sdata	slave	configurable	read data
SdataAccept	slave	1 bit	slave accepts write data
Sresp	slave	2 bits	transfer data

2.4 Target Hardware Implementation

ZeBu UF as shown in Figure 4 is an ultra-fast emulator provide environment for System-on-Chip (SoC) debugging and embedded software validation and has been used for the target hardware implementation. ZeBu Compiler is a specific tool design for emulation on ZeBu UF-4 board. It has comprehensive hardware debugging that give full visibility of the design and it also integrated popular simulation tools which are VC, NCS and ModelSim. Some of its debug features are static probes, dynamic probes, flexible probes and waveform generation. ZeBu UF system also provides in-circuit emulation (ICE) which allows simulation of DUT through actual hardware environment.

ZeBu UF-4 board has four FPGA devices based on Xilinx Virtex-4 LX200 that is equivalent of 6 million of ASIC gates on a single PCI card. It uses PCI to directly interface to a desktop PC running Linux OS. It has 512 MB of DRAM and 64 MB of SSRAM. The board comes with full suite of software tools including compilation and run-time software package. Table 2 and Table 4 show the board detail and its performance for different type of operations while Table 3 shows logic resources available in Virtex 4 LX200 FPGA device. The ZeBu compilation process is incremental and the Xilinx ISE place and route phase is parallelized to reduce turnaround time. ZeBu emulation platform is as accurate as running zero delay gate-level simulation compared with RTL simulation.

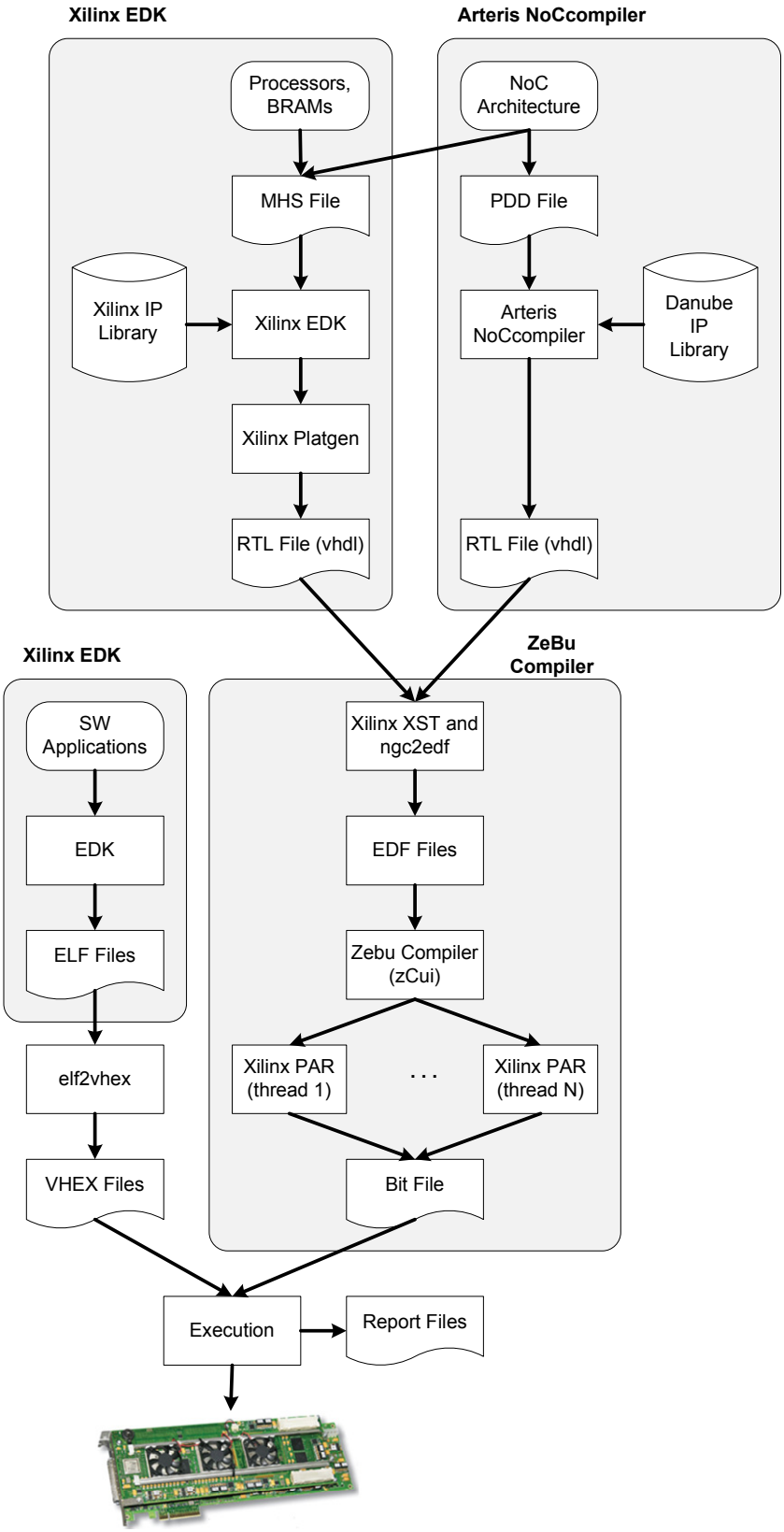


Figure 3: Design flow and EDA tools integration

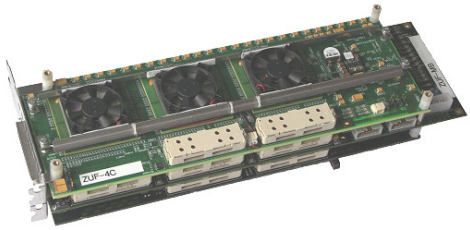


Figure 4: ZeBu UF-4 emulation board

Table 2: ZeBu UF-4 emulation board detail

Modules	Descriptions
FPGA	4 Virtex-4 LX200
DRAM	512 Mbytes
SSRAM	64 Mbytes
ICE	Smart and Direct

Table 3: Logic resources in Virtex 4 LX200

Modules	Descriptions
Slices	89088
BRAM (18 kb)	336
DSP	96

Table 4: ZeBu UF4 operating mode and performance

Operating Modes	Performance Range
Maximum capacity in ASIC gates	6 Million
Co-emulation with commercial HDL simulator	5 kHz-100 kHz
Co-emulation with signal-level C / C++ / SystemC	100 kHz-500 kHz
Co-emulation with transaction-level C / C++ / SystemC / SystemVerilog	500 kHz-20 MHz
Test vectors	100 kHz-500 kHz
Emulation with synthesizable testbench	<=20 MHz
In-circuit emulation, connected to target system	<=20 MHz
Emulation with SW debuggers via JTAG interface	<=20 MHz

2.5 MPSoC Architecture

2.5.1 Processor Architecture

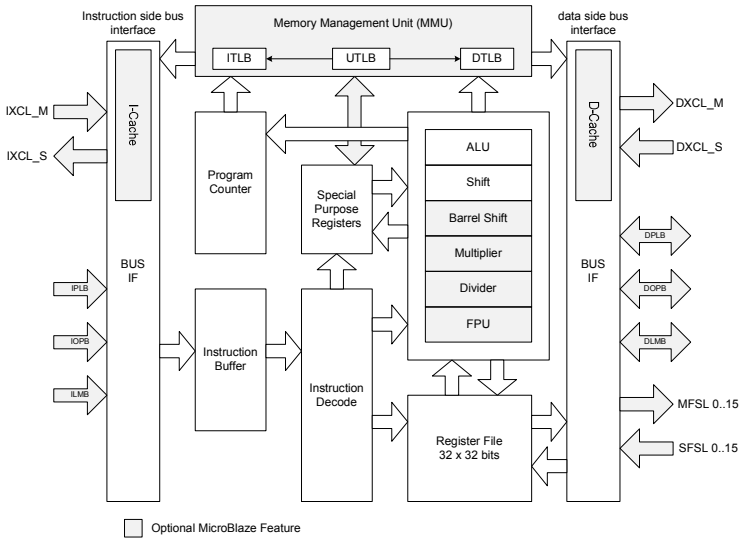
The MPSoC architecture is shown in Figure 5. All the MPSoC have been realized using Embedded Design Kit (EDK) tools from Xilinx. MicroBlaze softcore processor is used as masters and Block RAM (Random Access Memory) is used as slaves for all MPSoC designs. The MicroBlaze is an embedded soft core processors from Xilinx based on 32 bits Reduced Instruction Set Computer (RISC). It is highly reconfigurable and offer design flexibility in which users can select several configuration options such as floating point units, integer multiplier and integer divider. The MicroBlaze processor is configured to its full configuration and is given additional local 32 KB BRAM memory connected via two LMB BRAM Memory Controllers using two LMBs (Local Memory Bus) to provide Instruction Memory (via ILMB port) and Data Memory (via DLMB port). The MicroBlaze uses Fast Simplex Link (FSL) for its interconnection. FSL bus is a uni-directional connection, provides simple and fast point-to-point communication between two components in the EDK environment.

Since the NoC architecture has OCP-IP interface, therefore FSL to OCP-IP interface was developed. From NoC architecture to slaves BRAM, OCP to BRAM interface was also developed. Each MicroBlaze and BRAM has its own interface for FSL-to-OCP and OCP-to-BRAM. This makes the interface IP reusable for other different MPSoC NoC architecture. The Slave is composed of a BRAM block with a controller. It should be noted that one MicroBlaze (number 0) is connected to a timer via On-Chip Peripheral (OPB) Bus to allow onboard software performance monitoring as shown in Figure 5 (c).

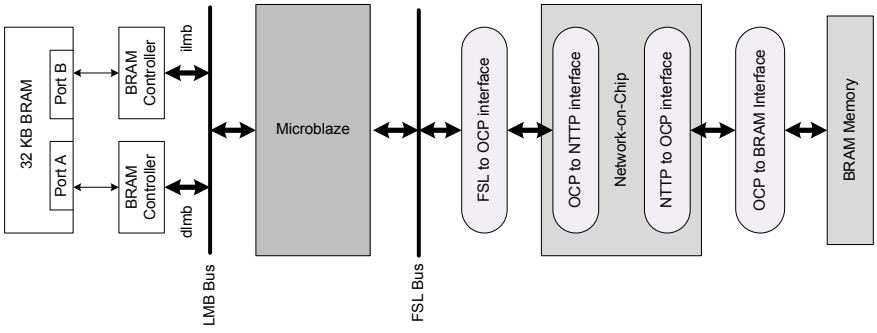
2.5.2 NoC Architecture

NoC topology is a geometrical configuration used to connect different network components. Many NoC topologies exist from a simple crossbar to the complex cubes. In this work, we used 2-ary 4-tree NoC topology for the MPSoC implementation. The topology is designed using NoCcompiler tool that could save much time rather than design it manually. The block diagram of 2-ary 4-tree with 16 masters and 16 slaves is shown in Figure 6. It has 56 switches in seven stages interconnection. Two masters are connected to one switch for the NoC input and two slaves are connected to one switch for the output. All the switches have two input and two output ports. The MicroBlaze processors and BRAM slaves are connected to the NoC through two interfaces; FSL-to-

OCP custom IP interface which will be described in the next section and OCP-to-NTTP interface which is included in the NoC through the NIUs elements within the NoCcompiler environment. Each Microblaze has its own FSL-to-OCP interface, which means that additional MicroBlaze processors can be included in the design with only modification of NoC architecture. The slaves address is identified based on the address specify in the NoCcompiler. The maximum address of 32 bit can be used for the slave address.



(a)



(b)

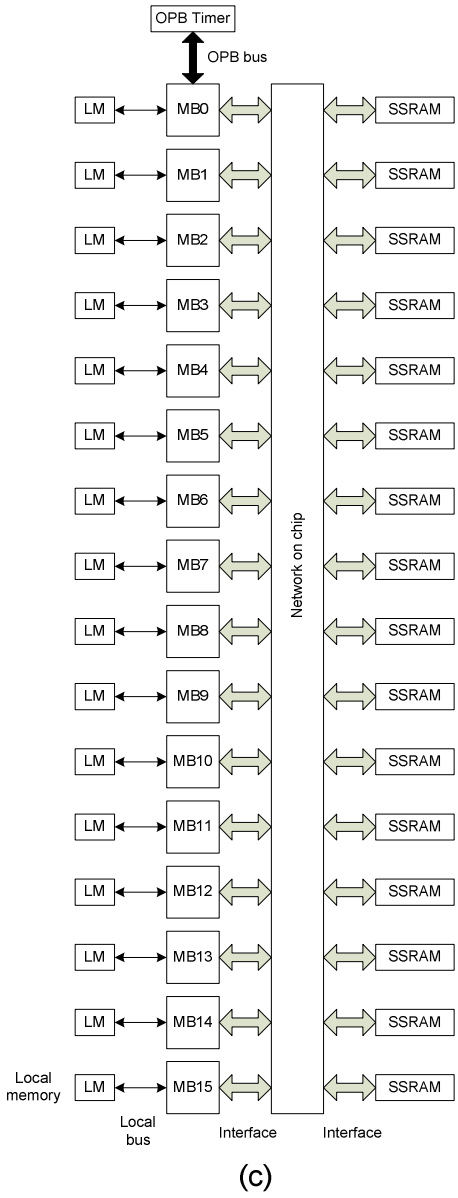


Figure 5: MPSoC with NoC architecture (a) MicroBlaze core block diagram [31] (b) interfaces between components for one MicroBlaze processor (c) complete block diagram of the system

2.6 Application: Discrete Cosine Transform

Discrete Cosine Transform (DCT) is a lossy compression technique, first introduced by Ahmed [32] which has been developed via the Discrete Fourier Transform (DFT). DCT has many advantages compare with other compression techniques and therefore it is employed in the international standard such as JPEG, MPEG, H.261, H.263, and DOLBY. DCT has been used in many digital image and video processing applications due to its advantages over other compression methods [33] [34]. An example of DCT application in JPEG standard is shown in Figure 7.

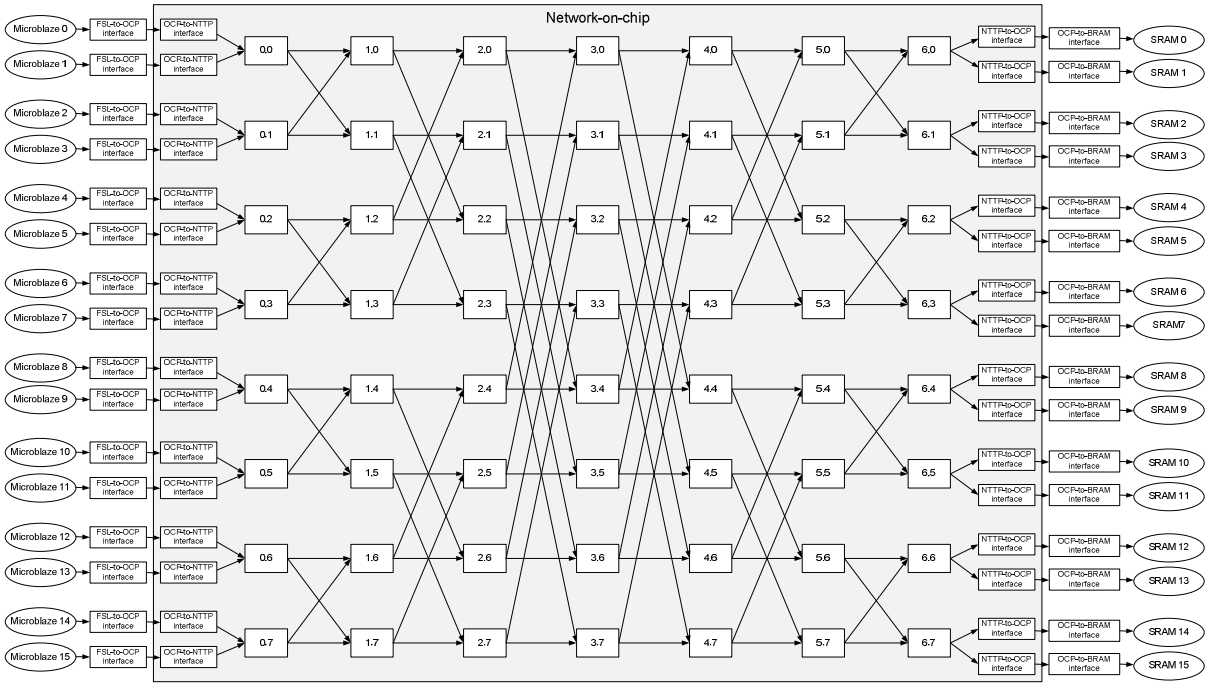


Figure 6: 2D NoC-based MPSoC architecture with masters and slaves connection

JPEG convened in 1987 under International Organization of Standards (ISO) has been developed the standard for still picture coding algorithms. It defined the standard for still images or pictures that uses two dimensional (2-D) 8 x 8 blocks DCT for transformation. There are four types of DCT; DCT-I, DCT-II, DCT-III and DCT-IV. These types are different in terms of their basis functions but all are still orthogonal transforms, meaning that the inverse transform is just reverse of the forward transform. Among them DCT-II is the most popular and widely used. For that reason, only this of type of DCT will be explained here. The 2-D forward DCT-II is given by the equation below [32];

$$X(u,v) = \frac{1}{4}C(u)C(v)\sum_{x=0}^{7}\sum_{y=0}^{7} p(x,y)\cos\frac{(2x+1)\pi u}{16}\cos\frac{(2y+1)\pi v}{16} \quad (1)$$

From equation (1), $C(0) = \frac{1}{\sqrt{2}}$, $C(k) = 1\ for\ k = 1,2,..7$ and $p(x,y)$= input image. In order to get the matrix form of the equation, the following equation is derived from the 2-D DCT equation in (1):

$$T_{i,j} = \begin{cases} \frac{1}{N} & if\ i = 0 \\ \sqrt{\frac{2}{N}}\cos\left[\frac{(2j+1)i\pi}{2N}\right] & if\ i > 0 \end{cases} \quad (2)$$

From equation (2), the DCT coefficient in matrix form, *T* is as following;

0.3536	0.3536	0.3536	0.3536	0.3536	0.3536	0.3536	0.3536
0.4904	0.4157	0.2778	0.0975	0.0975	0.2778	0.4157	0.4904
0.4619	0.1913	0.1913	0.4619	0.4619	0.1913	0.1913	0.4619
0.4257	0.0975	0.4904	0.2778	0.2778	0.4904	0.0975	0.4157
0.3536	0.3536	0.3536	0.3536	0.3536	0.3536	0.3536	0.3536
0.2778	0.4904	0.0975	0.4157	0.4157	0.0975	0.4904	0.2778
0.1913	0.4619	0.4619	0.1913	0.1913	0.4619	0.4619	0.1913
0.0975	0.2778	0.4157	0.4904	0.4904	0.4157	0.2778	0.0975

Therefore, to apply DCT on an image, a process which is based on matrix multiplication is performed as following:

$$DCT\ of\ image,\ D = TMT' \tag{3}$$

From equation (3), *T* is the DCT coefficient, *M* is the image data and *T'* is the DCT coefficient transposition.

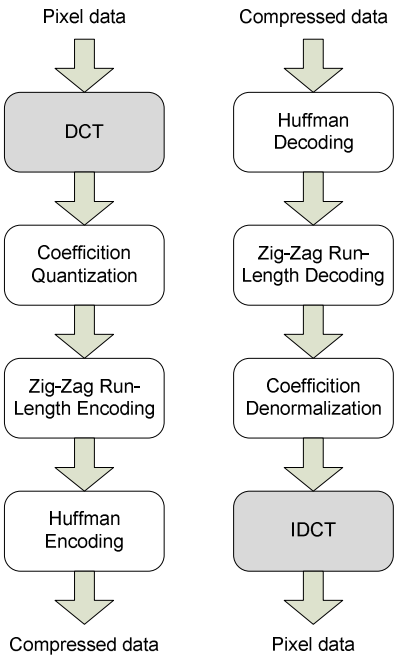


Figure 7: Example of DCT application in JPEG image compression standard

2.7 FPGA Implementation

The logic utilization after place and route stage of the MPSoC with NoC architecture is shown in Table 5. NoC uses much more logic resources than the other component in the MPSoC. This large amount of logic resources is due to the switches of the NoC. For this design, all MicroBlazes is not using all optional configuration as describe before such as floating point unit, integer multiplier and etc. from the figure, MicroBlaze consume larger FPGA space than other modules. NoC architecture also has high percentage of logic resources due to the large number of switches.

2.7.1 ENSTA APIs

Several Application Programming Interfaces (APIs) have been developed to ease the performance evaluation of the design which is PCI_function.c, ddr.h and synchro.h. The first function is used for function such as print value to the host PC. The second function is used for DDR interface on the board while the third function is used for synchronization of MicroBlazes for parallel processing. The synchro API is developed by manipulating the FSL bus connected to each MicroBlaze based on the OCP master command, *MCmd* as shown in Table 6. These APIs help to reduce design time as well as during the evaluation phase. Moreover, synchro API for example is independent of MPSoC architecture and thus can be reused in other design.

Table 5: Post place and route logic utilization for MicroBlaze with basic and enhanced configuration

Components	Logic utilization (%)	
	MicroBlaze with basic configuration	MicroBlaze with enhanced configuration
MicroBlaze	29.09	54.06
NoC	22.26	22.26
FSL-to-OCP interface	9.14	9.14
OCP-to-Memory interface	1.15	1.15
System	75.77	91.59

Table 6: OCP master command signals

MCmd	Command	Request Type
000	*Idle*	(none)
001	*Write*	write
010	*Read*	read
011	*ReadEx*	read
100	*ReadLinked*	read
101	*WriteNonPost*	write
110	*WriteConditional*	write
111	*Broadcast*	write

2.7.2 Parallel Programmming

In order to validate the design, we applied some functional application of parallel programming, in particular DCT application, which is based on matrix multiplication. Two matrix multiplication of 8 x 8 is performed to compute DCT of the sub-block image. An image which is 256 x 256 pixels is used as the input. This image is divided into 256 blocks of 16 x 16 pixels, where it is processed by dedicated MicroBlaze processor as shown in Figure 8 (mb0 is referred to MicroBlaze number 0). The image is stored in DDR memory on Alpha-Data board and each MicroBlaze will access this memory to get it image data to be processed. Each MicroBlaze processor has a unique identification number that can be used to perform parallel processing. For example, in this DCT application, the specific sub-block is calculated by a certain MicroBlaze based on its ID. In addition, to manipulate the parallel processing application for different number of processors, this ID can also be exploited. From the 16 x 16 pixels of the image, each MicroBlaze will have to perform DCT computation four times, on 8 x 8 pixels. In this evaluation, we use data parallel processing for DCT where each sub-blocks of the image is treated by the corresponding processor and the data is independent to other sub block.

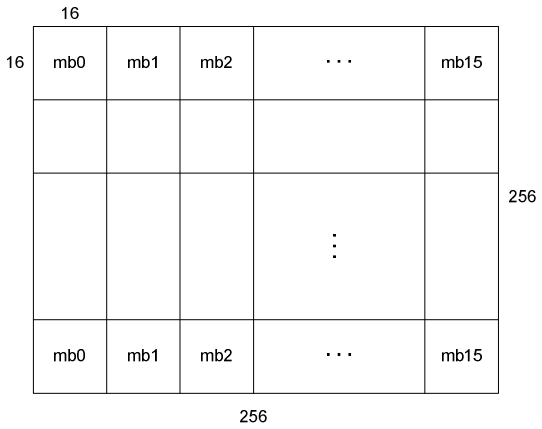


Figure 8: Processor allocation for data parallel of DCT application on 256 x 256 pixels image

In order to measure the execution cycles of the computation using more than one processor, locked synchronization is used based on the OCP interface signal, *MCmd*. For locked synchronization, it uses *ReadEx* signal followed with the Write signal. *ReadEx*, stand for read exclusive command sets a lock to the memory location after finished read the value. Then *Write* command clears the lock so that it can be read by other processors. Before calculation is started, first processors, MicroBlaze 0 will write start flag on certain memory location. This start flag will be read by other MicroBlaze processor and will start the computation of specific sub-block only if the start flag is set. Otherwise, the MicroBlaze will wait and continuously reading the value. For all Microblazes, they will read a specific sub-block from a memory and then perform DCT computation. The result will be writing back to the memory. The first MicroBlaze must wait until all 15 MicroBlazes finish the calculation and write a finish flag to a specific memory. After that the timer value is read at the first MicroBlaze. Only one synchronization process is used in this design which is before and at the end of the computation for all processors.

2.8 Results and Discussion

Several configurations of the MPSoC are designed to tested parallel programming application which are basic and enhance configuration. For basic configuration, the MicroBlaze has no hardware block while for the enhanced configuration, all hardware blocks available in the MicroBlaze processor such as barrel shifter, floating point unit, hardware divider and hardware multiplier has been used. In both MPSoC architectures, the NoC architecture has been kept the same. For any MPSoC configuration, the image data is stored in the DDR memory of the FPGA board and all processors will access it to perform DCT computation.

The execution cycle for different number of processors is shown in Figure 9 and Figure 10 for both MPSoC architectures with basic and enhanced configuration respectively while the detailed cycle number is shown in Table 7. The graph in Figure 11 shows that for both MicroBlaze configurations, similar trends can be seen where the execution cycle is decreased as we increased the number of processing cores. However, when we calculate the speedup for both MPSoC implementations, the speedup is reduced when moving towards larger numer of processor. For example in 16 processors MPSoC implementation, the speedup of 15.84 and 11.63 was achieved to compute the DCT application for MicroBlaze with basic and enhanced configuration respectively. As depicted in Figure 12, it can be seen that the speedup of MicroBlaze with enhanced configuration is less than MicroBlaze with basic configuration as the number of processor increased due to the face that

increasing number of processing cores cause the increased of logic utilization which in turn force the synthesizer to map to more than single FPGA device in the ZeBu UF4 board. When mapping the logic into multiple FPGA, the speedup reduction is mainly come from the inter-FPGA delay compared with single FPGA implementation which uses on-chip wires only.

Table 7: Execution cycles for different number of processors and different MicroBlaze configurations

Number of processors	MicroBlaze with basic configuration	MicroBlaze with enhanced configuration
1	1,047,702,312	28,809,664
2	524,190,774	14,435,031
4	262,567,070	7,291,651
8	131,577,991	3,886,736
16	66,122,127	2,476,443

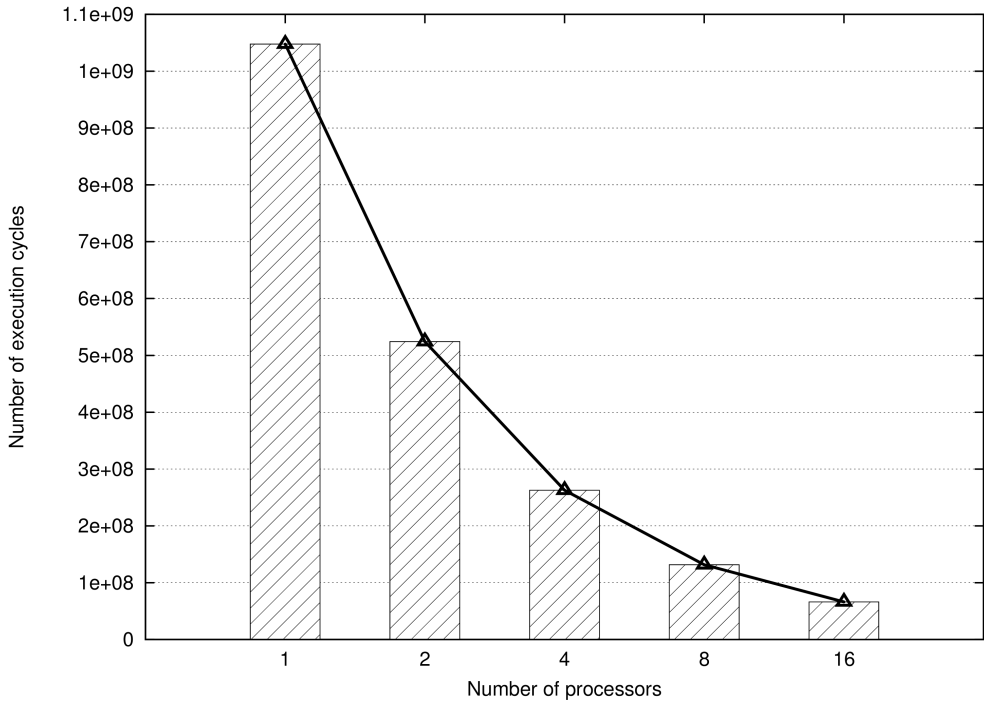


Figure 9: Execution cycles of MicroBlaze with basic configuration

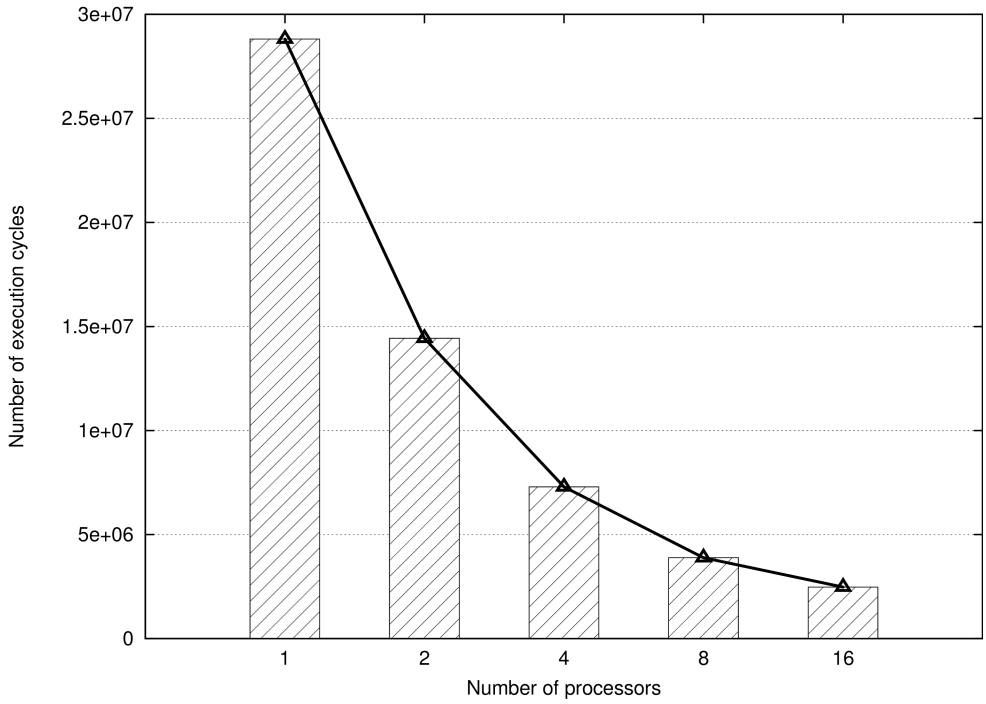


Figure 10: Execution cycles of MicroBlaze with enhanced configuration

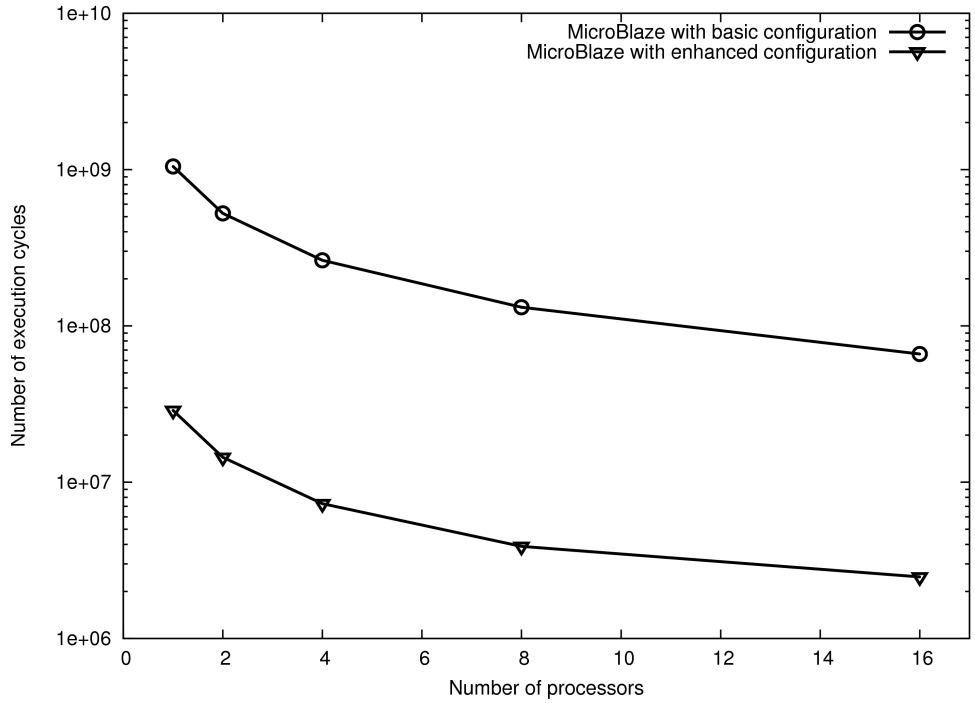


Figure 11: Comparison of execution cycles for MicroBlaze with basic and enhanced configuration

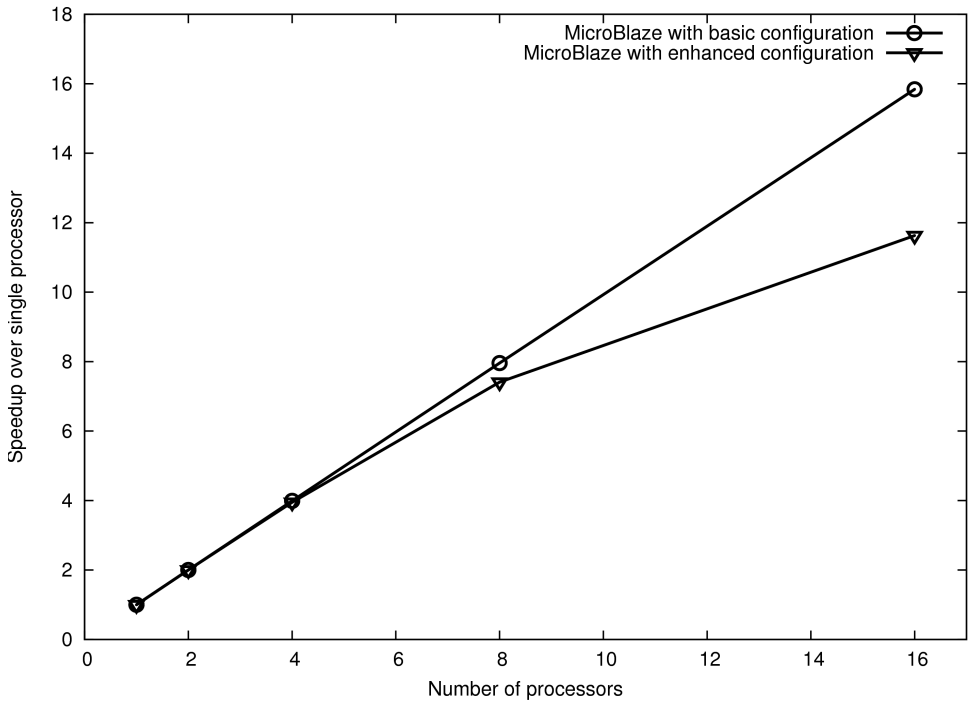


Figure 12: Comparison of speedup between MicroBlaze with basic configuration and MicroBlaze with enhanced configuration

Conclusion

In this chapter we presented MPSoC design and implementation with NoC architecture. 16 masters of MicroBlaze and 16 slaves of BRAM was designed and implemented on FPGA. The masters and slaves are interconnected using 2-ary 4-tree NoC topology. Parallel programming was used to evaluate the design. DCT application for processing 256 x 256 pixels image is tested on the design and the execution cycles are measured for different numbers of processors. Different configurations of MicroBlaze are also evaluated for the parallel programming application in order to get the impact on the hardware area as well as execution cycles. Results suggested that, as expected, MicroBlaze with enhanced configuration has lower execution cycles than MicroBlaze with basic configuration from a single processor to 16 processors. However, because the number of logic utilization is increased, the speed-up of MicroBlaze with enhanced configuration is worse than MicroBlaze with basic configuration especially when the number of processors is high.

CHAPTER 3
OVERVIEW OF 3D IC TECHNOLOGY

3D IC technology is the new approach to increase the device performance through stacking multiple dies or wafers on top of each other. It is seen as a viable solution now as technology node is approaching very deep sub-micron technology where many issues that are not a problem in the older technology become the critical parameters that must be taken into consideration due to the effects to the performance. In this chapter, we discuss various aspects of 3D technology including bonding methods, stacking techniques, TSV architecture, TSV material and manufacturing proceses, stacking orientation, standards in 3D technology and challenges faces by the technology. We first briefly describe some of the issues in 2D architecture the industry is facing as the industry is in the transistion to the 20 nm technology and beyond.

3.1 2D Architecture and Its Issues

As of today, 2D designs have been matured enough in many aspects including design tools, industrial supply chain, manufacturing equipments and packaging methods. The progress in 2D architecture is primarily driven by the performance improvement achieved by reducing transistor dimension within a period of time. By reducing transistor size, its switching speed is increased because of the short distance from source to drain, which essentially improve the overall speed of the designs. However, as the transistor size is getting smaller, reliability of the devices is significantly affecting its performance improvement trends. Issues arise for the advanced process technology are discussed as follows.

Small performance improvement is observed for 45 nm and it is getting smaller moving to 32 nm, 22 nm and so on because of several factors concerning effects of small transistor size such as gate delay. As the transistor become smaller, it allows higher device density, the performance is slowly increased because of increasing total delay (sum of gate and interconnect wire delay). Moreover, power consumption is also increased dramatically due to the transistor leakage increase and the transistor parameter's variation is also worsening due to small dimension [5]. A part from that, fabrication process is becoming more difficult and costly because of additional manufacturing processes such as double patterning methods. The number of lithography mask is dramatically increased when moving to the new process technology resulting exponentially increased of fabrication cost where it is estimated that the cost for a set of mask for 45 nm is $1 million while for 32 nm is $2 million. Transition to 450 mm wafer size could reduce the production cost but needs

substantial investment for the new manufacturing facilities as well as new equipments and the new extreme violate (EUV) lithography tools is progressing slowly which is mandatory for 10 nm technology and below.

Another issue is long interconnect wire due to increasing number of metal layers (up to 9 metal layers in 45 nm technology) as a result of higher transistor density which eventually leads to increasing design complexity [35] [36]. Evolution towards dense interconnect structure when moving to the advanced process technologies as shown in Figure 13 increases the number of buffers or repeaters along the interconnect wires substantially (Figure 15 (a)) in order to achieve the small delay requirements which can lead to via blockage problem and severely affects the interconnect wiring efficiency [37]. As technology generation scales down, interconnect delay is becoming higher than gate delay as plotted in Figure 14 (a) which have been retrieved from ITRS Interconnect Report [38], describing that interconnect delay trend is increasing while delay of NMOS is slowly decreased. For power consumption, long interconnects of global wires contribute to higher power because of close relationship with the number of buffers along the interconnect wires as well as high capacitance as shown in Figure 14 (b). Scaling transistor supply voltage (V_{dd}) is also progressing slowly that makes controlling power consumption even more difficult.

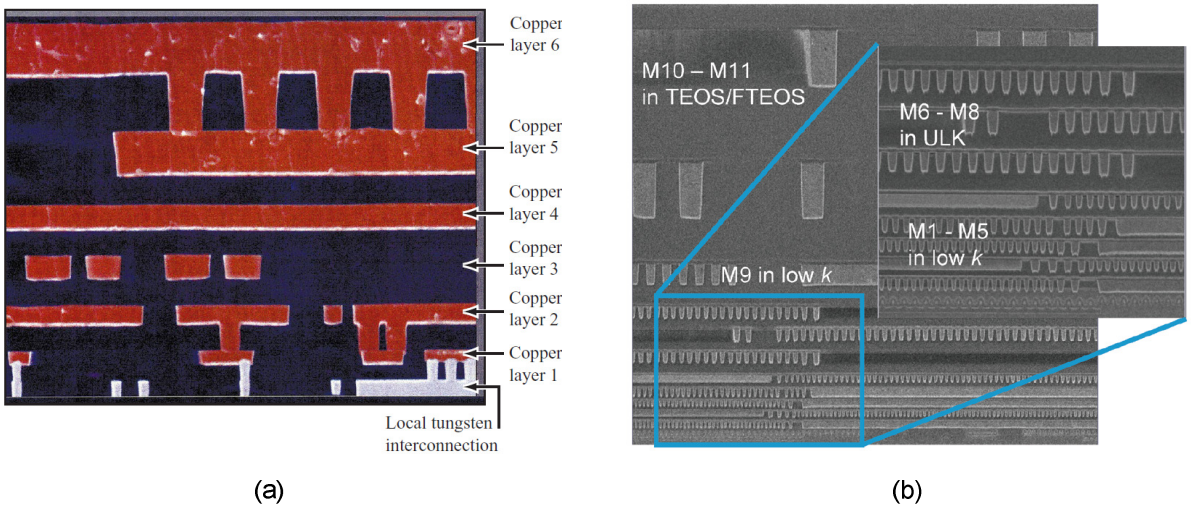


Figure 13: Evolution of the interconnection architecture for high performance CMOS logic (a) CMOS 7S process in 0.2 µm [39] (b) 45 nm process technology [40]

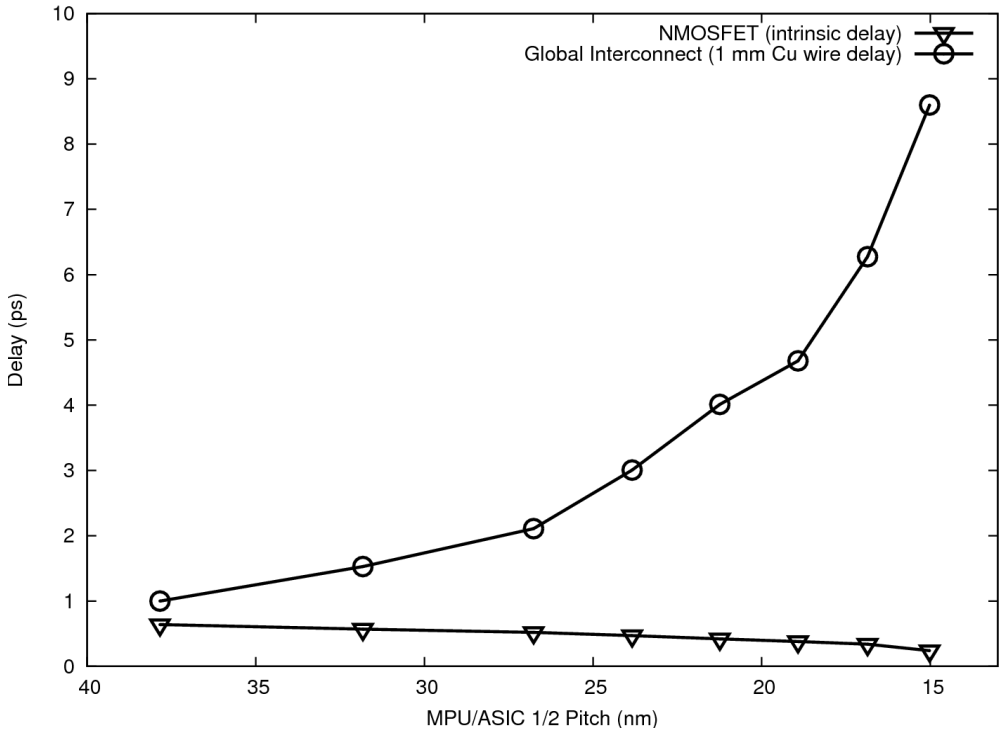


Figure 14: Interconnect and gate delay trends as technology node shrinking

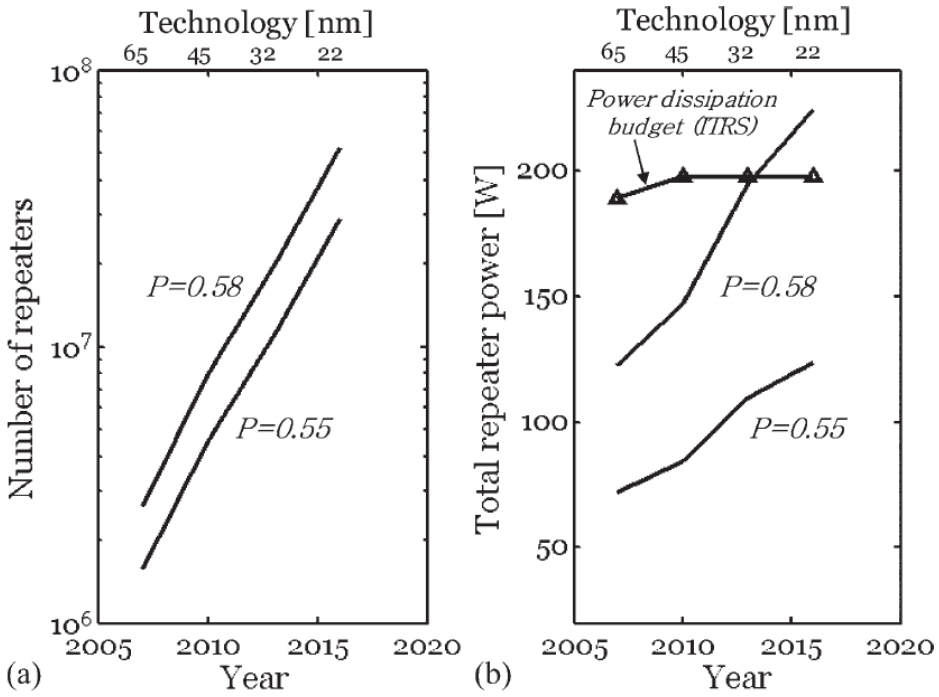


Figure 15: Technology scaling effects on (a) number of repeaters (b) total repeater power [41] where P is Rent's coefficient

These imitations of 2D architecture motivate designers to adopt 3D technology for several niches applications especially in mobile devices. 3D technology provides high intergration density with low cost solution for meeting the demand of smaller, cheaper and faster consumer electronic devices in the future.

3.2 3D IC Technology

3D integration offers less difficult method to achieve higher transistor integration for current application requirement compared with scaling transistor to smaller technology nodes. Rather than defining new process and researching solution for new challenges created by new process technology, 3D integration allows us to use older and matured technology for example 65 nm to double the transistor density and thus provide more functionality. Although 3D integration has been studied since 1986 in [42] , it was only in the research stage without commercialization due to the technology limitation at that time. With today's semiconductor technologies, 3D technology is feasible to be designed and implemented at a relatively low cost. However, in order for this technology to be widely adopted in the industry, several critical obstacles need to be solved which will be discussed later in this chapter.

Several means can be used to provide electrical connections between dies in the 3D stacking such as using TSV, wire bonding or contactless as shown in Figure 16. Stacking methods determine the trade off in terms of interconnection density and cost. Wire bonding technique provides chip to chip connection from outside the chip using bonding wires which is normally found in the 3D packaging technology. The chips are stacked using packaging technology glueing each other. This method is limited to the resolution of wire bonders, for example 35 μm for a 15 μm wire. This makes it more difficult when the number of I/O for each chip in the stack increases [43]. This type of structure has been demonstrated using more than two dies [44]. For contactless inter-die communication in 3D stacking, either capacitive [45] or inductive circuit [46] can be used to provide wireless communication between dies in the stack. Capacitive coupling communication is limited to a few micrometers while inductive coupling can provide longer distance [47].

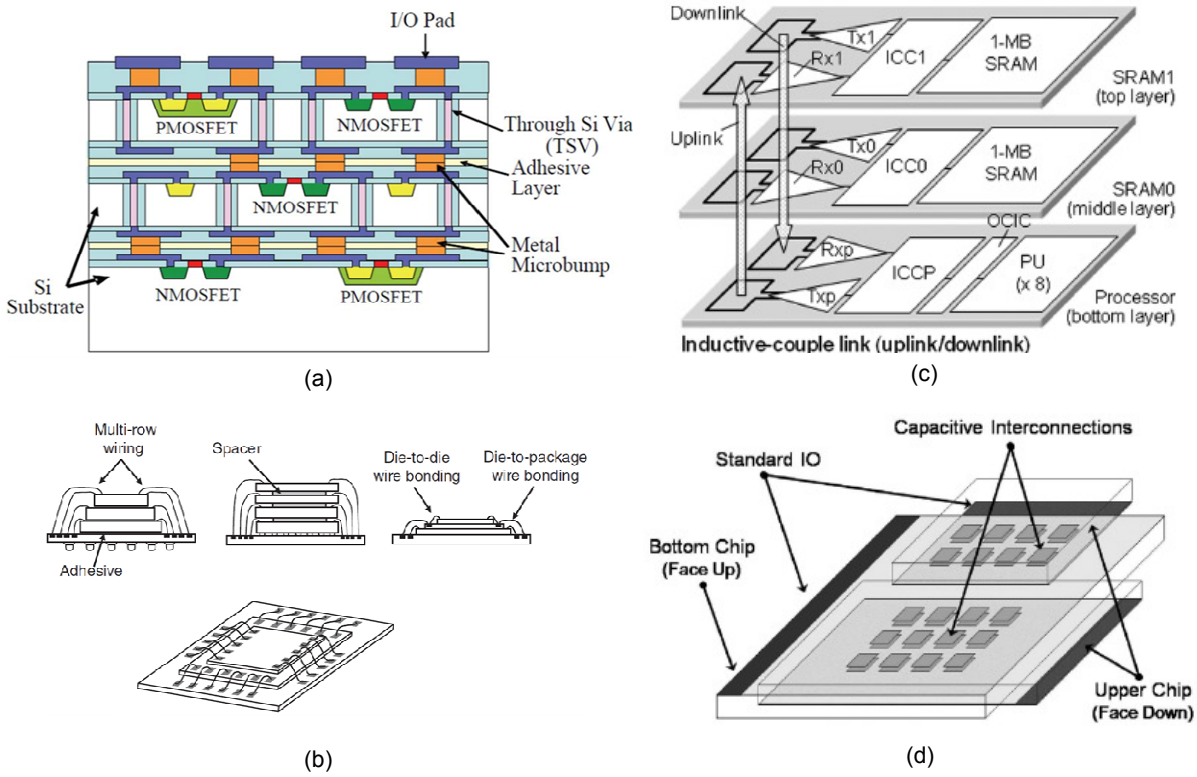
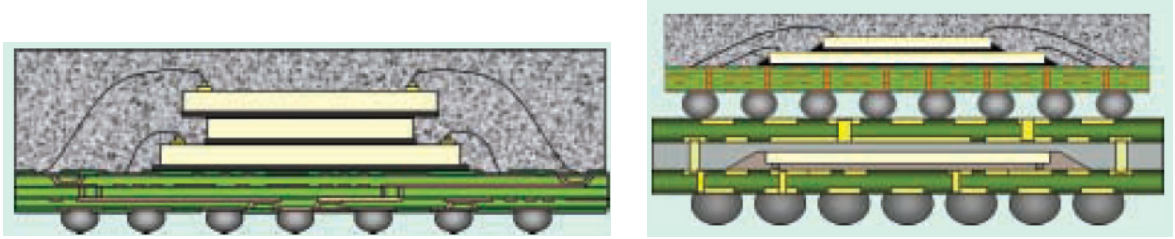


Figure 16: Different type of stacking methods (a) TSV [48] (b) wire bonding [49] (c) contactless using inductive coupling [46] (d) contactless using capacitive coupling [45]

3D technology can be categorized into 3D packaging, 3D IC using TSV and Monolithic 3D technology. In terms of packaging, 3D packaging technology uses wire bonding or flip-chip method for inter-die connection which has limitation on the number of connection it can offers due to the relatively large wire bond dimension as explained earlier. Some widely used type of this technology is system in package (SiP) [50] and package on package (PoP) as shown in Figure 17. This technology is currently being used in many mobile devices due to the high density integration in small form factor with low cost solution and it will be the mainstream technology for a few years in the near future until the solutions is found concerning the technical difficulties as well as high manufacturing cost of 3D IC technology using TSV for high volume production.



(a) (b)

Figure 17: Packaging types (a) System-in-Package (SiP) (b) Package-on-Package (PoP)

Monolithic 3D technology is currently still infancy but offers the highest vertical interconnection density. In monolithic 3D IC integration, instead of stacking wafers or dies, this technique build another device layer on top of a base device layer with proper isolation by means of sequential fabrications processes in the single wafer as shown in Figure 18 (b) where for example an inverter gate has NMOS transistor in the top tier while PMOS transistor in the bottom tier for 2 tiers design. Another variation of 3D Monolithic integration is that gates are stacked on top of other gates as shown in Figure 18 (c) called gate level monolithic integration. The first device layer can be fabricated using conventional process flow (bulk CMOS or SOI process) while the upper layers requires different methods such as laser crystrallization, seed crystallization and epitaxial growth [51]. This research is currently at an early stage although several papers have been published presenting the successful 3D integration for simple architectures such as inverter [52] [53].

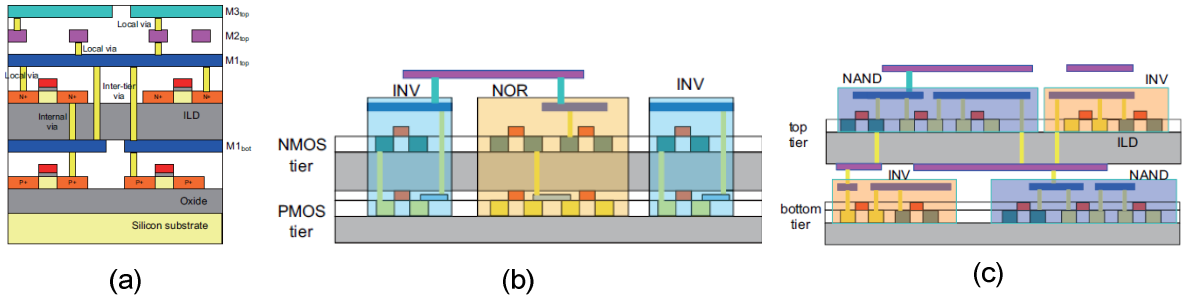


Figure 18: (a) Monolithic 3D IC complete structure (b) transistor level monolithic (c) gate level monolithic [54]

3D IC using TSV is an approach where dies or wafers are stacked and TSV is used for their electrical interconnections in a single package. Compared with 3D packaging, this technology offers higher interconnections density between dies and smaller structure due to vertical connection located inside the die area. The discussion in this chapter is mainly based on this type of 3D technology. There is also another form of this type of technology known as 2.5D where several dies are placed on top of a silicon interposer (active or passive interposer) which consists of several interconnect metal layers with TSVs that formed the connection between dies and to the outside interface. It provides very high device and interconnection density between dies and thus able to serve high memory bandwidth appealing for high performance mobile multimedia application [55] [56] [57].

3.2.1 Advantages of 3D IC Technology

One of the primary advantage of 3D technology is that the long interconnect wire length is reduced due to the stacking. The conceptual diagram of the interconnect wire reduction is illustrated in Figure 19. By stacking wafers, the maximum interconnect wire length can be reduced significantly depending on the number of stack as showed experimentally in [58] where reduction on average total wire length is more than 28% when stacking two to five wafers and from 31% reduction for the longest wire for International Symposium of Physical Design (ISPD'98) circuit benchmarks.

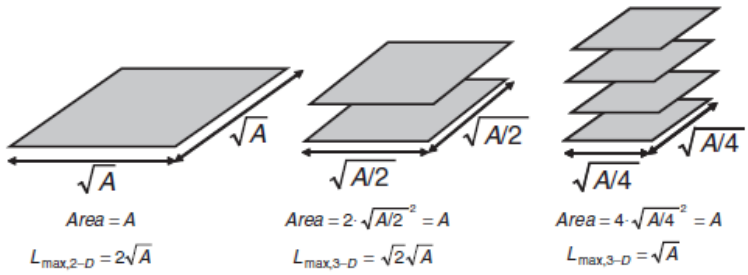


Figure 19: Reduction of wire length from 2D architecture to 3D architecture with different stacking levels [49]

An experiment on 3D FFT architecture [59] and for a microprocessor architecture [60] have proved the speed improvement of 3D design. As for 3D integration, reduction in global interconnect wires results of reduction of routing congestion and eventually increase the performance. Partitioning strategy has strong influence to the latency improvement and its scalability when stacking multiple layers [61].

The TSV delay should be considered when measuring its performance improvement and experimental results show that TSV delay is between 35 ps to 135 ps [62] and 16 ps for 20 µm height, which is lower than in wire delay in 2D architecture for example 219 ps of 2500 µm wire length for 4.5 GHz speed. TSV delay depends on several parameters such as diameter, height, pitch and its material. The height and pitch of the TSV affect largely its delay while the TSV diameter has the opposite effect [63]. The TSV resistance has less pronounced effect to the delay than TSV capacitance [64]. In terms of technology, SOI-based 3D technology will have less TSV delay than bulk-based CMOS technology due to smaller TSV dimension thus reducing its RC delay [61]. As we include more layers in the 3D structure, higher TSV delay will be noticed due to increase of TSV count.

The existence of vertical interconnection allows many opportunities of design optimization which cannot be accomplished using 2D architecture space. Wide variety of partitioning optimization can be explored in different number of stacking levels for hardware specific implementation in order to optimize the performance [59]. Apart from that, using automated tool for partitioning the design into 3D architecture could also provides considerable performance improvement in contrast of using manually optimized partitioning method.

Reduction of power consumption in 3D architecture is reported in [60] through experimental study on Kogge-Stone adder which showed reduction of power consumption around 8%, 15% and 22% for 2, 3, and 4 stacks over 2D architecture [60]. Furthermore, 3D integration is not only outperformed in terms of performance, it is also scalable as the design become more complex as, for example improvement of power consumption around 11%, 21% and 46% for 12, 36 and 72 bit Kogge-Stone adder [65].

3D architecture can also be used to alleviate memory wall problems by providing excessive and short vertical connections and also enables higher on-chip memory capacity needed especially by large multicore architecture [66]. Wide I/O architecture is another approach to mitigate memory wall problems by offering high data bandwidth through larger number of I/O pins for memory access [67].

Design miniaturization enabled by the 3D integration allows high density integration. For example, the chip footprint is reduced by 44% for four layers stack compared with two layers stack for 65 nm technology [68]. In another work, with seven layers stack, the total thickness is around 900 µm and less than 1 mm for 10 wafers stack. Therefore 32 GB non-volatile memory will become 320 GB memory in total [6].

Integration of heterogeneous technology is also less complex than in 2D architecture. This heterogeneous technology integration allows different architectures for example analog, RF, sensor, memory to be integrated without difficult fabrication process as each architecture is produced using their own optimal process technology and then they are integrated in 3D structure with specific techniques such as wafer bonding. Besides, heterogenous logic architecture can also be implemented using different process technology such as 95 nm for processor with 65 nm memory as demonstrated in [69]. This enables SoC applications with better capability for meeting embedded system requirements such as real time processing and lower power consumption and also support for future SoC design that is highly heterogeneous structure [70].

From the SoC perspective which has digital and analog blocks integrated in a die, 3D integration also overcome noise isolation problem in 2D mixed signal architecture because of digital components tend to be error prone affected the nearby analog/RF circuitry in the same chip. Noise isolation can easily be formed in 3D architecture by separating the analog/RF and digital components in different silicon layers [71].

3.2.2 TSV Technology

TSV is a method that uses via across different layers of active silicon. Material used for TSV are Tungsten (W) [72], Copper (Cu) [73] [74] and Poly-Silicon (Poly-Si) [48]. Poly-Si material is stable and has less effect on device characteristic than other materials. However, Copper or Tungsten is more suitable for the TSV due to lower resistance. Copper is most commonly used because it has good thermal conductivity compared to Tungsten and Poly-Si. However, as will be discussed later in 3D challenges and issues, Copper TSV create stress effect due to large difference of coefficient of thermal expansion (CTE) between Silicon substrate and Copper, which is not the case for Tungsten TSV. A detailed comparison of via filling material can be found in [75]. Tungsten has longer delay compared with Copper TSV for any diameter size and therefore is only used in the research [76].

TSV formation process consists of several steps which are drilling, insulation, filling or metallization, FOEL formation, BEOL formation, handling attachment, wafer thinning and backside processing. The order depends on the TSV formation techniques either via-last, via-middle or via-first. As handling thin wafer is a great challenge, it can be avoided by thinning the wafer after bonding with another thick wafer. This could also prevent yield reduction because of additional processes for wafer handling during bonding and debonding. TSV can be manfactured either using DRIE or laser drilling process where the TSV structure using both process is shown in Figure 20. Using DRIE, also known as Bosch Process, is a widely used method that can produce high aspect ratio but at the expense of higher cost compared with laser drilling method. Laser drilling method limits the TSV diameter to about 10 μm. Additionally, it is a serial process and thus does not suitable for designs with high TSV count [77].

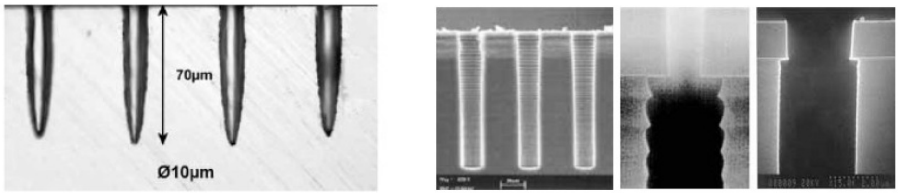


Figure 20: TSV manufacturing using (a) laser drilling process and (b) DRIE process [77]

TSV allows high interconnection density between stacked chips. For example 120,000 interconnections for 12.5 mm^2 area of 3D chip containing processor and memory [78]. Another reported work achieve 10^3 interconnections for W TSV with 10 µm TSV pitch in the area of 1 mm^2 [79]. Another important thing is TSV lining or TSV insulation in order to insulate from the Silicon substrate. Most commonly used material is Silicon Oxide which can be deposited using Chemical Vapor Deposition (CVD) or Atomic Layer Deposition (ALD).

There are different techniques to implement TSV for stacking multiple tiers such as via-first, via-middle or via-last. Each method poses different characteristics. In via-first technique, TSV is formed before the BEOL structure. Therefore we have a relatively small size compared to the other two methods [80]. Via-middle approach creates the TSV after BEOL and before FEOL formation. The size of TSV is in between the size of via-first and via-last methods. While in via-last approach the TSV is formed after the BEOL formation which results a large TSV size. This method also poses great challenge during TSV formation process formation in order to prevent damage to the devices that have already been formed. All the TSV formation methods are illustrated in Figure 21.

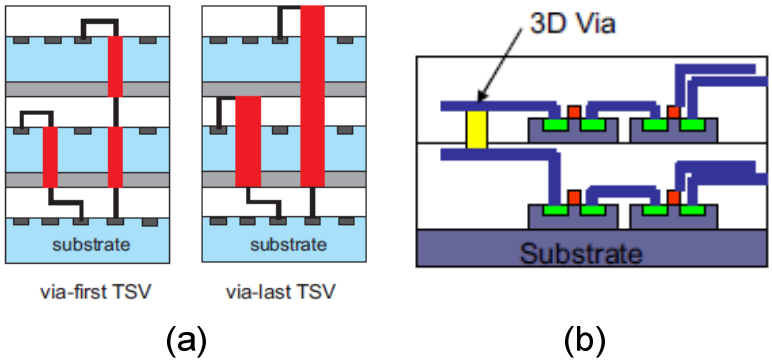


Figure 21: TSV stacking methods (a) via-first and via-last in bulk CMOS (b) via-first TSV in SOI CMOS [81]

3.2.3 Stacking Techniques for 3D IC Technology Manufacturing

Stacking methods can be implemented in several ways such as wafer to wafer, die to wafer or die to die as summarized in Figure 22. Wafer to wafer method is mostly used for 3D integration due to the low cost than the other two methods. However, it suffer from low yield due to the bonding bad yield dies compared with other methods which support known good dies to be bonded. Another downside of wafer to wafer stacking is that it is limited to the dies with same size in the wafers making it provides high production throughput. Die to die methods has a drawback of high cost due to the bonding of each die but can be used to bond different die sizes.

Wafer to wafer stacking is not necessarily achieving good 3D integration because of the difficulty ensuring stacking with known good dies (KGD). If there is not enough sufficient testing of the dies, the possibility of stacking wafers that contains bad dies exist and it will affect the reliability. However, die to wafer stacking or die to die stacking provides better control and opportunity for achieving good stacking. This is because known good dies has been ensured before integration with a wafer for 3D integration.

	Chip to chip	*Chip to wafer*	*Wafer to wafer*
Pros	Flexible, use of KGD	Flexible, use of KGD	Low cost
Cons	Handling and bonding	Handling and bonding	Overall yield, chip size
Wafer thickness	<4 μm to >150 μm	<4 μm to >150 μm	<4 μm to >150 μm
Bonding technology	Solder Metal to metal Adhesive	Solder Metal to metal Adhesive	Solder or metal-oxide bonding Adhesive

Chip-to-chip bonding
C4 solder or microbumps

Chip-to-wafer bonding
Thin solder/intermetallic

Wafer-to-wafer bonding
Cu-to-Cu or oxide bonding

Figure 22: 3D stacking methods comparison [82]

From the point of bonding orientation, several methods exists such as face-to-face, face-to-back and back-to-back as shown in Figure 23. For two-tier implementation as in Tezzaron 3D technology, face-to-face orientation is the best way where inter-die connections use microbumps and thus does not block any routing layers. For more than 2 tiers as in MIT Lincoln Lab 3 tiers technology, both face-to-face and face-to-back orientation are used where all the inter-tier connections is done through TSV structure. For 2 tiers implementation, back-to-back orientation does not benefit the

performance because of TSV delay compared with face-to-face orientation using microbumps structure. However, there exist certain case where back-to-back connection is needed as demonstrated by [83] where 2 face-to-face connection is then connected through back-to-back orientation shown in Figure 24 to further increase the device layers which eventually increase its density and performance.

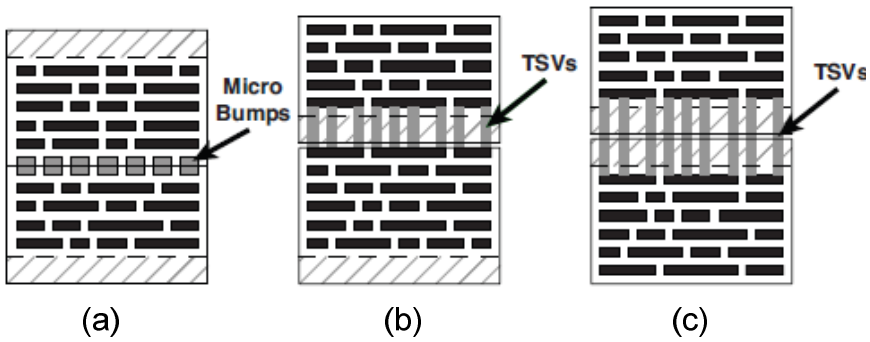
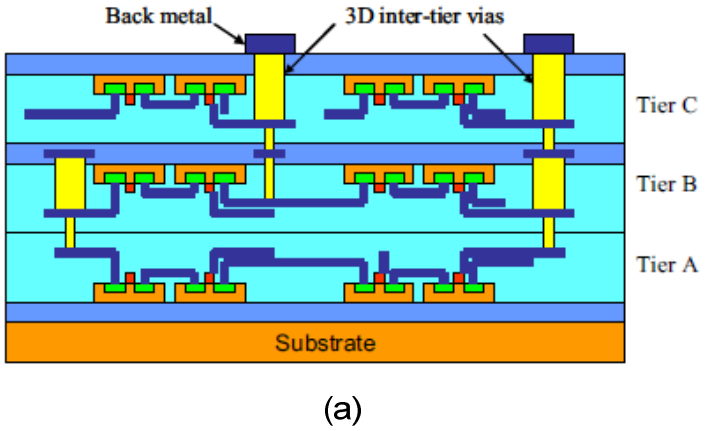


Figure 23: 3D stacking orientations (a) face-to-face (b) face-to-back (c) back-to-back



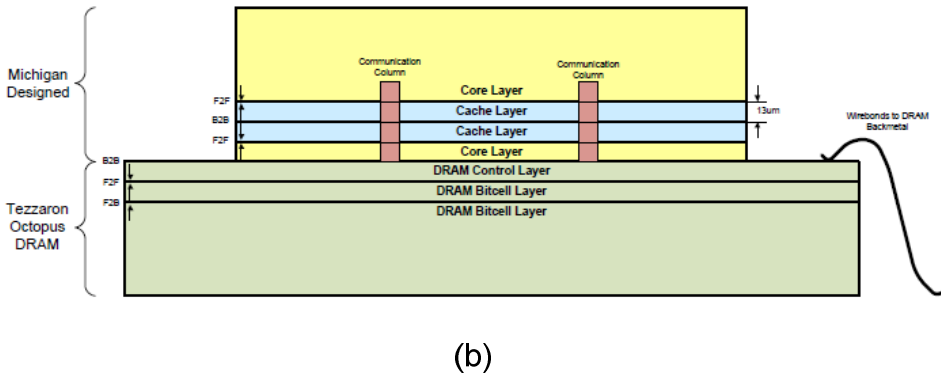


Figure 24: Examples of 3D stacking orientations (a) face-to-face and face-to-back using MIT LL technology [84] (b) face-to-face and back-to-back using Tezzaron Technology [83]

In terms of physical bonding implementation, several methods are available such as metal to metal bonding, direct oxide bonding and adhesve bonding. The comparison among these techniques is summarized in Table 8. Generally, metal to metal bonding is better because it gives mechanical and electrical connection between wafers and Copper is the most commonly used material. However, it suffers from high temperature processing for example more than 350ºC. Bonding alignment is the key parameter for achieving high interconnection density for metal to metal bonding [85]. This high temperature must be carefully monitored because it can damage the bottom layer and affect the device. Cu metal bonding has the strength of more than 50 MPa. Other material use for metal bonding is Gold (Au) and has less strength property which is around 10 MPa. However it can be processed at a lower temperature than Cu metal [86]. Au-Au bonding method has lower temperature process than Cu-Cu bonding. Cu-Cu bonding suffers from long process time and small throughput.

Adhesive bonding uses low temperature processing. However, there is possibility of contamination from the adhesive material to the devices. The material to be used generally should have excellent adhesion, high thermal and electrical resistance which is divided into photosensitive and non-photosensitive material. Among the used materials are Benzocyclobutene (BCB), Polyimide and Parylene. BCB is the most commonly used because it has the highest bonding strength which is more than 20 MPa [86].

Direct bonding uses oxide or silicon substrate for the bonding material [87]. It is done at room temperature followed by high temperature anneal to get a covalent Si-O-Si structure. Therefore, it has the highest bonding strength compared to other methods. The problem is that it is very sensitive to contamination for example a 1 μm diameter particle could create a 1 cm diameter void when bonding eight wafers [86].

Hybrid bonding techniques combining metal with adhesive bonding have also been reported [88] [89] [90]. Compared with direct bonding as metal to metal bonding, it has the advantage of bonding conformality due to the adhesive reflow during the bonding but need a high quality wafer-level CMP process for bonding interface and also has the potential problems with adhesive material such as moisture absorption and thermal stress due to the CTE mismatch [91].

Table 8: Comparison of wafer bonding technology [92]

Properties	Bonding types		
	Metal-to-metal bonding	Direct bonding	Adhesive bonding
Interconnect functionality	Mechanical and electrical	Mechanical	Mechanical
Bonding process requirements	Clean Oxide free Planar surface High temperature – 300ºC to 400ºC, 30 to 45 minutes	Need for very small roughness (nm) Clean Surface activation for low temperature anneal Room temperature prebond, anneal 300ºC, 30 minutes	Uniform coating Adhesive compatibility with post bond processing 150ºC to 320ºC, 10-20 minutes
Advantages	Direct electrical contact between stacks	Excellent alignment High throughput	Insensitive to particles and roughness
Disadvantages	Need for controlling thermal expansion at high processing temperature Alignment consideration	Extreme sensitive to particles and surface roughness	Little high temperature stability Weak mechanical rigidity Lowest bonding accuracy
Company/institute	Tezzaron	Ziptronix/MIT LL	IMEC/RPI

3.2.4 Partitioning Granularities for 3D Architecture Implementation

With the new vertical structure, designers have more design space to explore to find the best possible solution for 3D architecture implementation under the target performance constraints such as design time, die area and functionality. Table 9 presents the pros and and cons of different 3D stacking granularity that have been upated slighty from the table presented in [93] where it is clear that designing 3D architecture at very fine-grained partitioning (monolithic transistor level) will give the highest area, wirelength, power and performance benefits due to compact transistor footprints.

Table 9: Comparison of stacking granularities for 3D architecture design

Stacking granularity	Potential benefits	Design considerations
Entire cores, caches	Added functionality, more transistors, mixed-process integration	Low: Reuse existing design, can use existing 2D EDA tools
Functional unit blocks	Reduced latency and power of global routes provide simultaneous performance improvement with power reduction	Must re-floorplan and retime paths. Need 3D block-level place-and-route tools. Existing 2D blocks can be reused
Logic gates (block splitting)	Reduced latency/power of global, semi-global and local routes. Further area reduction due to the compact footprints of blocks and resizing opportunities	Need new 3D circuit designs, methodologies and layout tools. Reuse existing 2D standard cell libraries. Requires high number of vertical connections (limited by the size of TSV)
Transistors (monolithic)	Highest area, wire-length, power benefits due to compact transistor footprints.	Need new 3D standard cell libraries and require very high design effort. Need new 3D EDA tools

3.2.5 Tezzaron 3D IC Technology

Besang, Ziptronic and Tezzaron are the industry players that offer different kinds of 3D integration services while IMEC [94], Stanford University and MIT Lincoln Lab are among active research bodies from academic investigating 3D integration architecture. We will describe Tezzaron technology in details as the experiment conducted in this thesis is based on this technology but briefly explain about 3D integration from other company/institutes.

Ziptronic [95] technology offers 3D technology using covalent oxide bonding achieving high interconnection density. The advantages of this method compared with thermal metal bonding are that it allows lower temperature bonding process which is good for device reliability. A part from that, it is more cost efficient due to less complicated processes. They announce two technologies for bonding which are DBI and ZiBond.

MIT Lincoln Lab 3D integration technology [96] is based on 180 nm technology, three-metal layers, three-tiers stacking FDSOI wafer with additional top metal and back metal for bonding with another tiers. Wafer-to-wafer bonding is achieved using oxide bonding while inter-tier connection is through TSV. The first two wafers are stacked face-to-face while the third wafer is stacked face -to-back.

Besang [97] offers 3D integration by forming several single-crystalline silicon layers above a silicon substrate with metal interconnection between them using normal vias (not using TSV). High density memories using this type of process has been demonstrated in [98].

Tezzaron technology [99] is based on wafer level stacking as shown in Figure 25. The wafer is bonded using thermal metal bonding using Cu and Tungsten material [78]. Tezzaron has developed several TSV architectures and one of them is FaStack technology. They achieve alignment accuracy for the wafer around 0.5 μm.

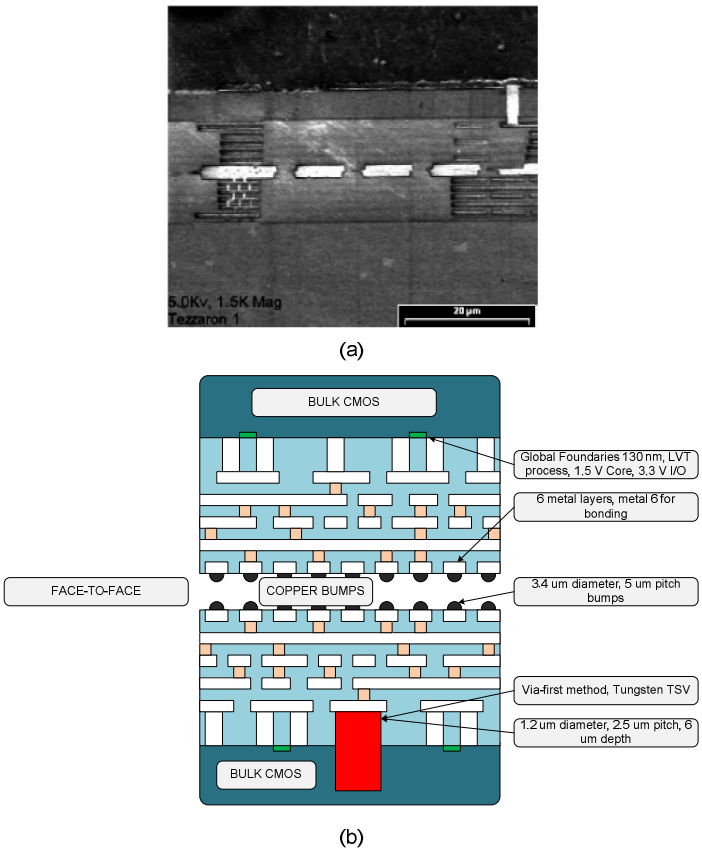
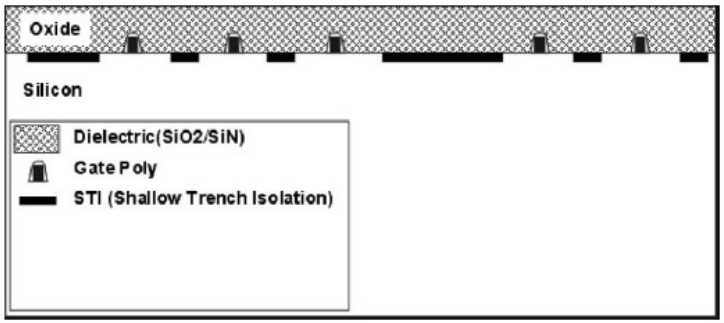
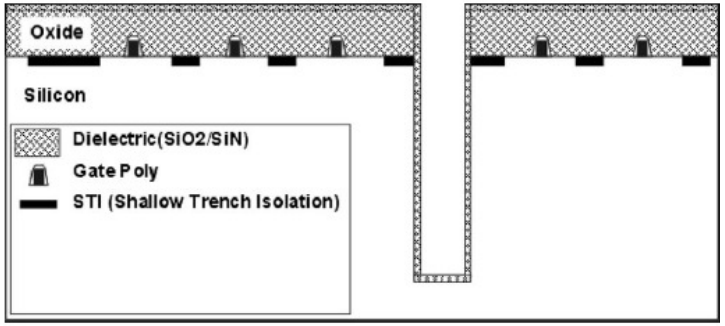


Figure 25: Two-tier Tezzaron 3D face-to-face stacking (a) cross section image of the manufactured device (b) cross section of the stacking technology with the corresponding parameters

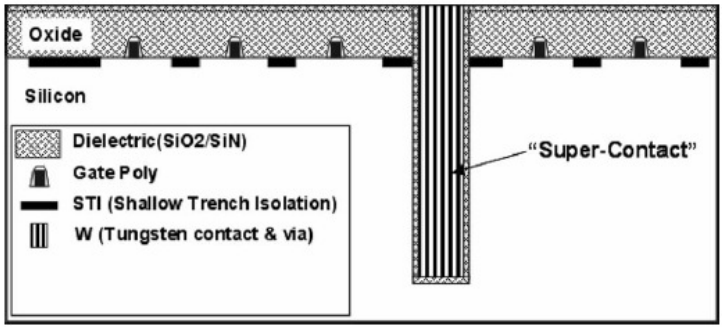
Tezzaron used via first methods face-to-face bonding followed by face-to-back stacking techniques for the three tiers implementation as shown in Figure 26. It has demonstrated several 3D test chip such as CMOS sensor, 3D FPGA, mixed signal ASIC and processor/memory stack. Additional wafer layer increased about 15 µm thickness making it possible for many more layers of stacking for high capacity architecture such as memory. Because the wafer is thinned after bonding, so there is no need for wafer handling process to help to reduce yield losses because of additional process of attachment and deattachment of wafer handle.


Oxide
Silicon
Dielectric(SiO2/SiN)
Gate Poly
STI (Shallow Trench Isolation)

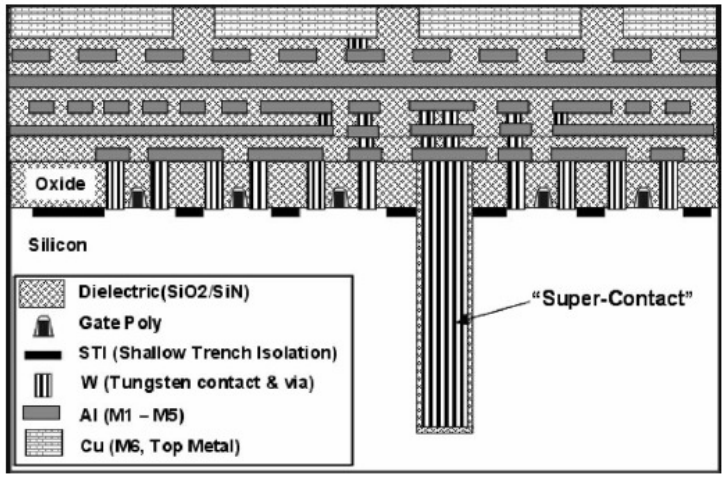
(a)


Oxide
Silicon
Dielectric(SiO2/SiN)
Gate Poly
STI (Shallow Trench Isolation)

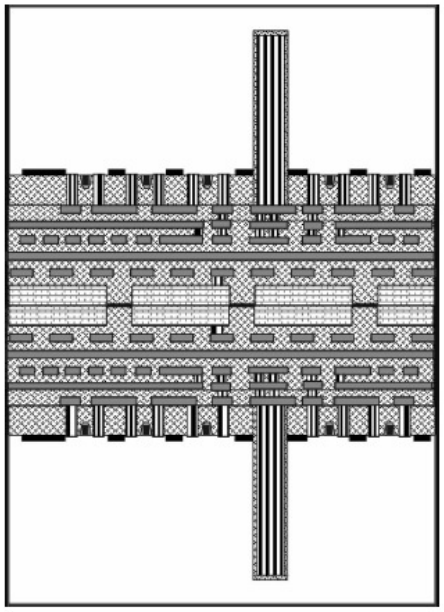
(b)


Oxide
Silicon
Dielectric(SiO2/SiN)
Gate Poly
STI (Shallow Trench Isolation)
W (Tungsten contact & via)
"Super-Contact"

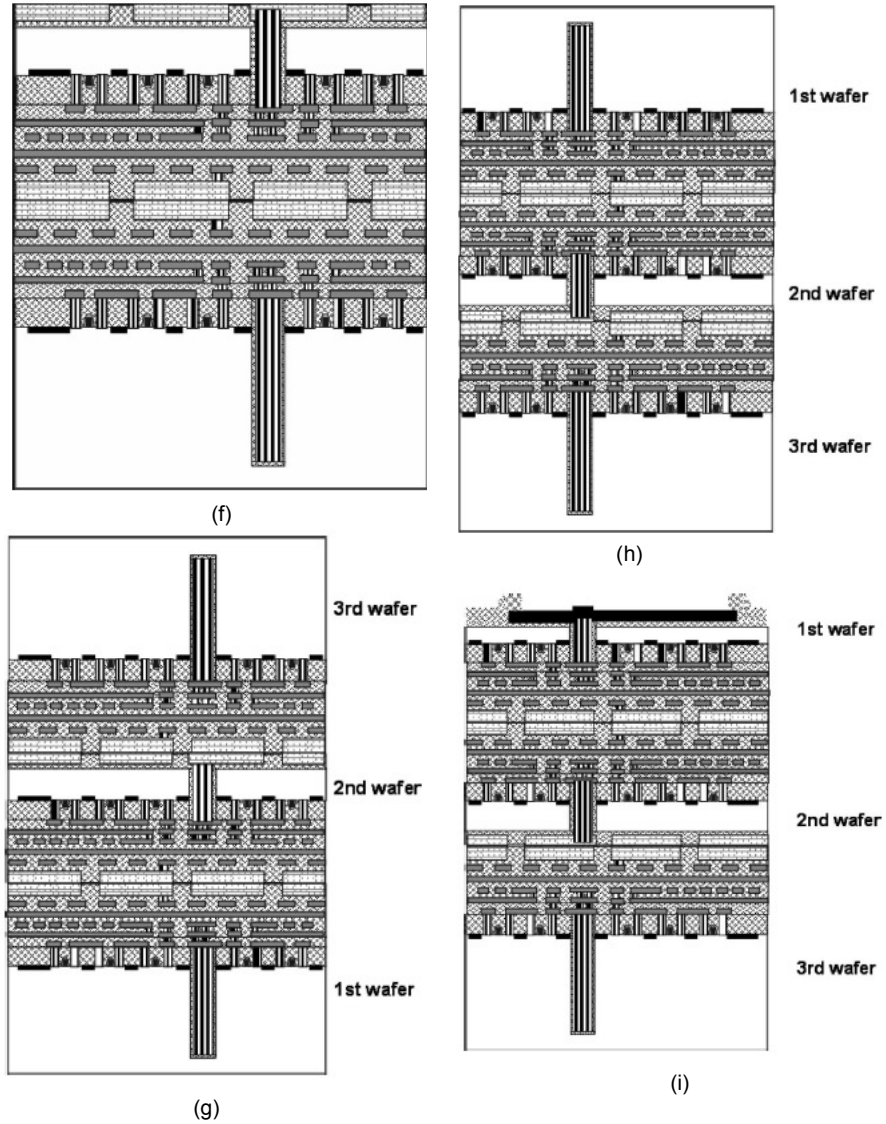
(c)


Oxide
Silicon
Dielectric(SiO2/SiN)
Gate Poly
STI (Shallow Trench Isolation)
W (Tungsten contact & via)
Al (M1 – M5)
Cu (M6, Top Metal)
"Super-Contact"

(d)



(e)



(a) Wafer cross section after transistors formation but before metal contact

(b) TSV etching with insulation using SiO_2/SiN

(c) TSV filling with Tungsten

(d) Interconnect metal layers formation with Copper must be the last metal layer

(e) Bonding process to the second wafer

(f) Thinning top wafer followed by Copper bonding pads formation

(g) Bonding the third wafer using the same process

(h) Inverting the complete stack for final processing

(i) Thinning the backside of first wafer followed by Aluminium layer deposition for wire bonding

Figure 26: Tezzaron 3D technology manufacturing process

3.3 CMOS Scaling vs 3D IC Technology

The electronic design community is now facing the consideration of either to choose to continue to scale the transistor size or to move to the 3D integration to design high performance architecture at the lowest possible cost. It is widely believe that 3D integration will not solve all the issues but it could provide alternatives for some specific application domains rather than using advanced technology nodes to achieve the same objectives such as high device density and higher performance using old process technology. As for CMOS transistor scaling, there exist several critical issues as follow:

1) Growing fabrication cost: non-recurring engineering (NRE) costs and lithography cost are increasing towards smaller feature size. In order to migrate to new technology node, the industry needs new manufacturing facilities such as new EUV lithography tool and this investment is increasing from one technology node to the other.
2) Significant effect of process variation: moving towards smaller transistor size, process and parameter variation is worsening. Small change in a parameter in the process will affect the product significantly. Process variation is becoming more important for the semiconductor process. Various new techniques are needed for mitigating the effect of process variation for 45 nm technology node compared with the previous technology [100]. For example, among scaling challenges beyond 32 nm technology are [101]:
    - Increased off-state current from degraded drain-induced barrier lowering drain induce leakage current (DIBL) and subthreshold slope (SS) by poorer short channel effects significantly limits the effective gate length shorten than approximately 15 nm.
    - decreasing oxide thickness, t_{ox} provides better channel control but with the penalty of increased gate leakage current and increased channel doping, eventually decreased mobility and increases random dopant fluctuations (RDF) and degrading minimum operating voltage.
    - Decreasing gate pitch increases the parasitic capacitance contribution for contact to gate and epitaxial layer to gate thus increasing overall gate capacitance.

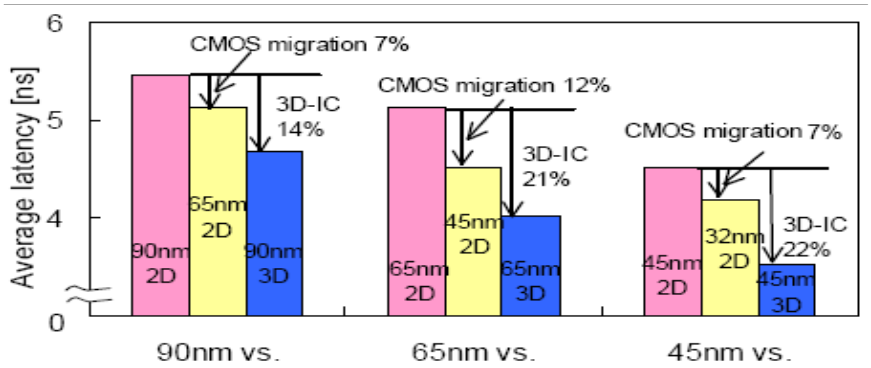


Figure 27: Performance improvement comparison of CMOS migration vs 3D integration [76]

On the other hand, 3D integration provides some solutions to continuing Moore's law as has been described in detail. In addition, experimental result in Figure 27 shows that performance improvement of average latency for multicore architecture is about double when designed using six stacks 3D technology for any technology node compared with migrating to smaller size. From this comparison, it shows that 3D integration is a promising solution for future VLSI direction.

For the manufacturing cost, compared with 2D architecture, 3D architecture reduces the cost to more than 25% for two stacks architecture for the case of die to wafer bonding [68]. Study of 3D design choices on manufacturing cost showed that single die integration is not cost effective for a system of more than 70 million gates because of the single die area become too large. On the other hand, 3D integration has lower cost due to the improve yield as a result of smaller die sizes [102]. 3D integration reduces cost even if the defect density is high. For example, for a defect density of 0.003/mm^2, two stack reduce cost by 46% while four stacks reduces cost by 61% [68].

3.4 Challenges of 3D IC Technology

Despite advantages it offers, there exist several challenges that need to be tackled for making 3D integration appliable in consumer electronic devices. Although several works have been reported in the literature, detailed study particularly on implementing real 3D chip is needed to drive the 3D integration technology. The challenges and proposed solutions are as follows.

The nature of 3D stacking causes heat that cannot be transferred out of the chip, particularly the heat generated far away from the heat sink. Unlike 2D architecture where heat generated for all components can be transferred directly to the heat sink through heat spreader because it can be placed just above the components. The important effect of stacking in 3D structure is increased peak temperature [103] [104] in the chip where it can reach more than 100ºC. Two things that are

very important as a result from this high temperature which is temperature variation and hotspot. Temperature variation between dies can be around 10ºC for two stacked dies [105]. Hotspot in the 3D chip can be up to more than 100ºC while temperature difference between stacks can be 1-20 ºC. Two things affecting the reliability of the chip are mean time to failure ratio (MTTR) and time to breakdown (TTBD).

The problem of thermal or high temperature in 3D chip is very serious. Heat is one of the factors affecting the reliability of the device. As the transistor size getting smaller, leakage power is becoming large, for example 25%-40% leakage of total power for 90 nm technology while 50%-70% leakage of total power for 65 nm technology [106]. Leakage power increases exponentially with temperature. Every 15ºC increase of temperature cause the interconnect delay variation around 10%-15%. Temperature increase is also causing electromigration which increases exponentially and eventually reduces product life time by four times [107]. Several methods have been proposed for thermal management techniques to solve thermal problem in 3D integration such as thermal herding which place the most frequently switch blocks near the heat sink [108] and using thermal vias to transfer heat out of the chip [109]. Thermal aware design focusing physical design stage such as floorplan and placement [110]. Thermal management techniques using dynamic frequency scaling (DFS) which proposed that dies near heat sink can be assigned using higher frequency (eventually higher temperature) while workload that has strong thermal influence is assigned to the die that has stronger cooling efficiencies [105].

Thermal stress is another effect of thermal problem when integrating using TSV. This is due to the different CTE property of Silicon, Cu, Silicon Dioxide and W as shown in Table 10. The CTE of Cu is larger than W when compare with Silicon. This means that Cu TSV has stronger stress impact on Silicon. However, W has lower thermal conductivity than Cu. Thermal stress cause timing variation around ±10% for an individual cell [111]. Thermal induced stress in 3D integration causes crack at the interface of TSV and Silicon substrate and between Cu interconnects and low-k insulator [112]. This effect is strongly influent on device reliability. The effect of thermal stress of Cu and W TSV is shown in Figure 28. Cu TSV produces high thermal stress up to 750 MPa. CTE mismatch causes stress to the Silicon substrate near the TSVs. Tensile stress will be created when Silicon is at room temperature because Cu electroplating and annealing process under high temperature [113].

Table 10: Electrical and thermal properties of several materials

Material	Electrical resistivity (nΩm at 20ºC) [58]	Electrical conductivity ($S \cdot m^{-1}$)	Thermal coefficient of expansion (10^{-6} m/K) [58]	Thermal conductivity (W/mK) [58]
Silicon (Si)	-	1.2×10^{-5}	4.68	149
Silicon dioxide (SiO_2)	-	-	Varies	10
Copper (Cu)	17	60.7×10^{6}	16.5	401
Tungsten (W)	53	18.2×10^{6}	4.5	173

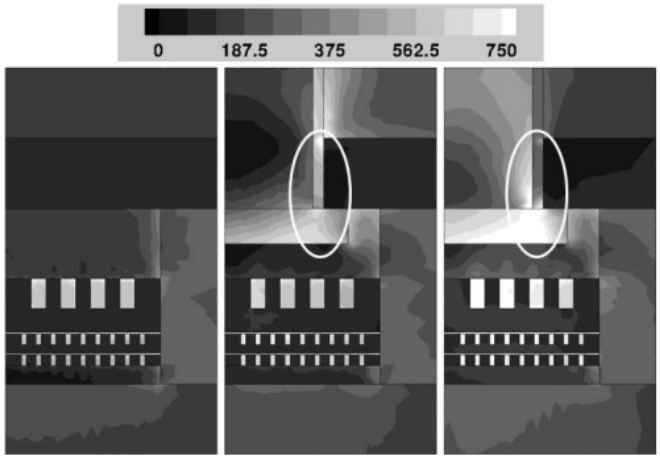


Figure 28: Thermal stress from Copper and Tungsen TSV material [114]

Power delivery network is important to make sure cells receive enough current supply for correct operation. When dies are stacked, the farthest die from the power supply source could potentially decrease the required voltage level for reliable operation and therefore IR drop measurement is crucial for 3D design. A part from that, TSV-related aspects such as TSV shape (cylindrical, square, coaxial) and size, TSV spacing, TSV topology have an impact to the power delivery quality in 3D architecture [115] [116] and therefore analysis at chip level (complete 3D) structure is required for accurate trade-off analysis.

Yield issue is also another important factor which has to be considered for 3D structure. 3D integration reduces overall yield due to the manufacturing process (TSV formation, bonding/debonding wafers) which will be lower as there is more stacks [117]. Yield loss of 3D integration can be due to the global planarization during wafer thinning process to stack them. New yield test is important to protect the performance of 3D integration structure [118]. Yield losses in 3D Integration are due to defects in the wafers during fabrication process or during 3D integration

process such as bonding steps. Wafers with large die area will have more possibility defects and thus will have lower yield compared with wafer with small die area. Yield is reduced when die size is increased and defect density increased [119]. Yield model has been developed for assisting in the decision making process about assembly trade off for 3D integration such as how many stacks to build and what is the optimum die size [119]. Techniques for yield improvement for 3D RAM design have also been proposed through inter-tier redundancy which is the design have spare blocks to be used when there are faulty memory blocks [90]. Some of the techniques for improving yield losses such as redundant resources such as memories and sensor arrays [78] and improving 3D integration process [120].

Other challenge for 3D integration is testing. With additional structure, TSV like testing is needed. TSV testing is a problem because probe size is large (35 μm) compared with the small TSV size such as 5 μm diameter with 10 μm pitch [121]. This is because normally this probe is used for testing 2D architecture which has wire bonding structure. Additionally, there are many challenges for testing 3D architecture such as test architecture, test access mechanism, test scheduling, test pattern, testing under thermal and power constraint which is important especially for testing at run time. New defects create during 3D integration process introduced new type of defects such as in TSV or bonding structure which require distinctive testing techniques. Testing for 3D architecture is a great challenge because functional units of processors at microarchitectural level can be partitioned at more than one layer. Testing is difficult because each layer does not have a complete functional system and thus require new testing strategy. Furthermore, pre-bond and post-bond testing is also vital to ensure only KGD dies is integrated in the 3D architecture and TSV formation as well as bonding structure do not have defects [122].

3.5 3D IC Technology Standards

Standards are the important thing in the electronic industry before any new technology can be widely adopted. The need for standards in 3D technology is crucial as each companies have their own perspective regarding this technology making it difficult for them to work together to establish business model for high volume production. Furthermore, 3D standards are also important for companies that provide manufacturing equipments and design tools before any investment can be made regarding what specific type of 3D technology will be adopted. Generally, the standards can be categorized in terms of manufacturing standards and designs standards and a number of working groups have been formed to start proposing 3D standards. In terms of manufacturing, several standards have been published related to wafer bonding, die stacking, TSV and reliability concern

such as JEP158: 3D Die Stack Reliability Interation by 3D JEDEC, MS1-0307: Guide to Specifying Wafer-Wafer Bonding Alignment Target and MS5-1211: Test Method for Wafer Bond Strength Measurements Using Micro-Chevron Test Structures by SEMI. On the other hand, 3D design standards include test architecture of the 3D design, design exchange formats and verification formats. Among the published standards are IMIS – Intimate Memory Interface specification by 3D-IC Alliance and JESD229: Wide I/O Single Data Rate (Wide I/O SDR) by JEDEC. As for now, there are many more standards to come as they are now being actively discussed by various standardization bodies.

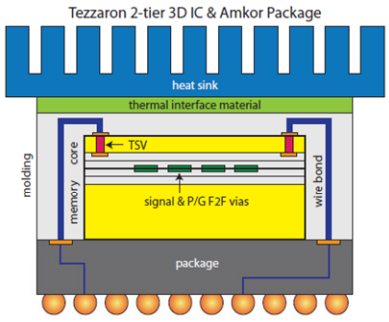
Table 11: Summary of published 3D standards

Category	Name	Organization	Status
Memory Interface	IMIS™ - Intimate Memory Interface Specification	3D-IC Alliance	Published
Guide	JEP158: 3D Die Stack Reliability Interaction	JEDEC	Published
Memory - Wide IO DRAM	JESD229: WIDE I/O Single Data Rate (WIDE I/O SDR)	JEDEC	Published
Guide	MS1-0812: Guide to Specifying Wafer-Wafer Bonding Alignment Target	SEMI	Published
Metrology	MS5-1211: Test Method for Wafer Bond Strength Measurements Using Micro-Chevron Test Structures	SEMI	Published

3.6 State of the art of 3D IC Architecture Implementations

We discuss several 3D chips that have been taped out using various 3D technologies available to date targeting at different objectives over the last few years in order to demonstrate the feasibility and benefits brought by the 3D IC technology. The purpose is to present a strong evidence regarding the capability of 3D technology to improve performance and thus making it a viable alternative solution for certain application domains. There are other 3D chips that have been fabricated without using TSV such as [123] and it is not discussed here.

In [124], they designed 64 cores using two-tier Tezzaron 3D technology and Global Foundries 130 nm standard cells as shown in Figure 29. The Tezzaron technology uses via first method with face-to-face bonding wafer level stacking. They created custom VLIW in-order processors in five stages pipeline architecture to have efficient power efficient inter-core communication by removing large and complex data structure. The project demonstrated large memory bandwidth of 3D stacking architecture which is up to 63 GB/s. Inter core communication is achieved using 4 buffers architecture in each core to their neighbouring cores. Global barrier was used for synchronization for cores. The design can be run at 277 MHz and has been tested with several parallel benchmarks proving the correct functionality. Each processor core has 1.5 KB instruction memory and 4 KB data memory. TSV architecture has 1.2 µm diameter, 5 µm pitch, 6 µm depth based on Tungsten TSV. Microbumps architecture has 3.4 µm diameter and 5 µm pitch. TSV is used for chip I/O interface and tier to tier connection is using microbumps. Each tier has 5 mm x 5 mm silicon area. A custom architecture is created modified from JTAG IEEE 1149.1 for off chip interface which are test control state machine, and by using four pair of tdi and tdo for each 4 blocks, 16 cores per block.



(a)

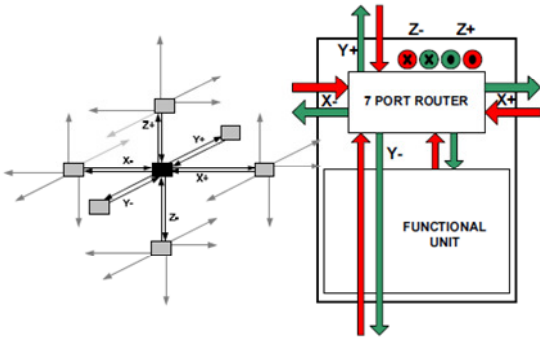
Process technology	Global Foundries 130nm
Die size	$5 \times 5mm$
Core footprint	$560 \times 560\mu m$
Core-to-core pitch	$570\mu m$
PG 3D connections/core	668
Total PG 3D connectiions	42,752
Data 3D connections/core	116
Total data 3D connections	7,424
TSVs/IO pad	204
Total IO TSVs	47,940
Dummy TSVs	6,540
Total maximum IR-drop	$78mV$
Maximum operating frequency	$277MHz$

(b)

Figure 29: 3D-MAPS (a) architecture and (b) design summary

In [84], they successfully demonstrated a working 3D mesh NoC in 3x3x3 configuration using via-last method from MIT LL 180 nm technology FDSOI process as shown in Figure 30. The 3D NoC has 2 mm x 2 mm die area per tier. The MIT Lincoln Lab has 3 metal layers for each tier, with a metal layer between two top tiers and a metal layer on top of the entire stack. Its TSV architecture has 2.5 µm x 2.5 µm with 3.9 µm pitch. The two bottom tiers are bonded face-to-face and the third tier is connected using face-to-back. The NoC used deterministic XYZ routing algorithm where each router port has two unidirectional links with 16 bit links. There is a functional unit connected to each router designed using linear feedback shift register (LFSR) to be able to send and receive packets. The design was routed with 145 MHz with the power consumption of 120.5 mW. The goal of the test chip is to validate the high level system simulator for 3D NoC they have developed. The

router used adaptive XYZ routing algorithm and it has no memory buffer and therefore each flit takes one cycle to travel across each router.



(a)

Freq	145 MHz
Gate count	~81,000
# of TSVs	576 (3 tiers)
Die size	2 mm x 2mm

(b)

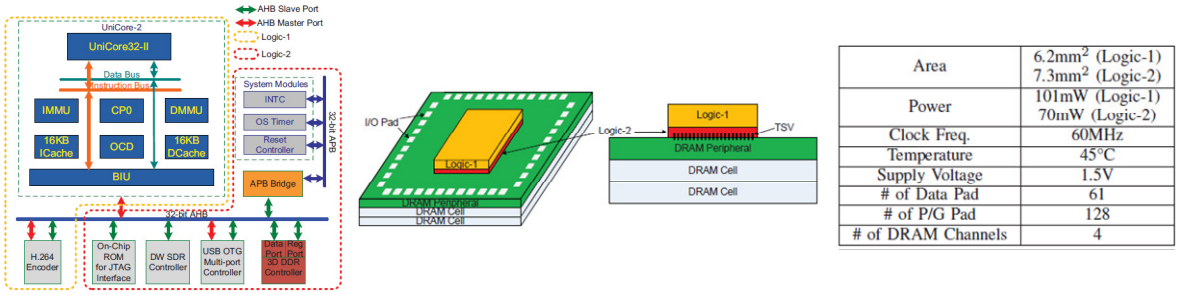
Figure 30: 3D NoC (a) architecture and (b) design summary

Another 3D chip implementation is 3D FFT processor of 1024-point memory-on-logic for synthetic aperture radar (SAR) using MITLL 180 nm FDSOI technology as shown in Figure 31 [59]. The FFT is radix-2 Cooley-Tukey FFT. The chip demonstrated the benefits of 3D technology showing 53% decrease in average wire length, 24% increase in maximum operating frequency and 25.3% reduce in the total silicon area using the customized design flow specifically for this implementation. The 3D die area is 23.40 mm^2, 4.8 mm x 4.8 mm and the design can be operated at 79.4 MHz (12.6 ns) with 409.2 mW power consumption. Block level partitioning has been used where processing elements and memory is placed in the three tiers such that memories are close the processing elements.

Figure 31: 3D FFT processor layout

In [125], they implemented two-tier logic of 2.5 mm x 5 mm die with a three layer 8-channel 3D DRAM stacked on top using Tezzaron 3D technology with Global Foundries 130 nm process technology as shown in Figure 32. The purpose of the work is to demonstrate the feasibility of 3D IC architecture for SoC design. The partitioning scheme is done manually at block level where USB controller, H.264 encoder block with its local memory is placed in top tier and other blocks in

bottom logic tier where AHB system bus connects between both logic tiers. The design run at 60 MHz and the DRAM can run at 133 MHz.

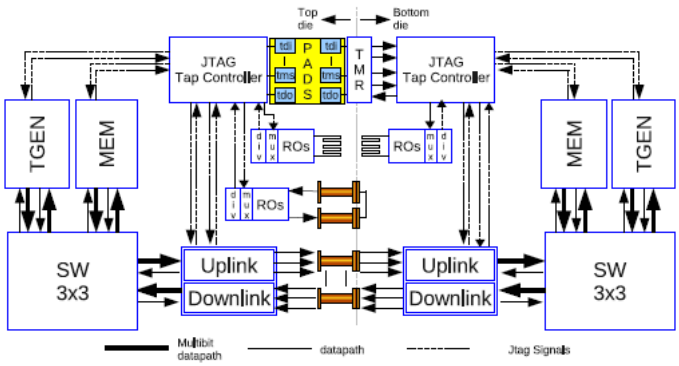


Area	6.2mm^2 (Logic-1) 7.3mm^2 (Logic-2)
Power	101mW (Logic-1) 70mW (Logic-2)
Clock Freq.	60MHz
Temperature	45°C
Supply Voltage	1.5V
# of Data Pad	61
# of P/G Pad	128
# of DRAM Channels	4

(a) (b) (c)

Figure 32: 3D SoC (a) architecture (b) 3D stacking diagram (c) design summary

In [126], they demonstrated the feasibility of 3D NoC design in two-tier 3D technology implemented using die-to-wafer bonding of IMEC 130 nm process technology with one poly and two metal layers as shown in Figure 33. The design has 1 mm^2 die area with 100 TSVs and 12 IO pads. The Copper TSV diameter is 5 µm, 25 µm depth and 10 µm pitch inserted after FEOL and before BEOL formation. Each tier has a traffic generator, a slave memory, a 3x3 switch and a JTAG controller with fault tolerant test structures. The traffic generator is programmed using JTAG controller which can send and receives flits from NoC. A slave memory is 64 bit arranged in 8 words wide 8 bit. Vertical links are unidirectional for the router and targeted for static faults like stuck at and stuck open fault. The design can run at 25 MHz at 0.4-1.5 voltage supply synchronously. Each vertical link was implemented using 2 TSVs for fault tolerant mechanism.



Technology	130 nm, 1 Poly, 2 Metal
# of Transistors	69,000
Die size	1 mm^2
Freq	100 MHz
# of TSVs	148
# of IOs	12

(a) (b)

Figure 33: 3D NoC with fault tolerant (a) architecture (b) design summary

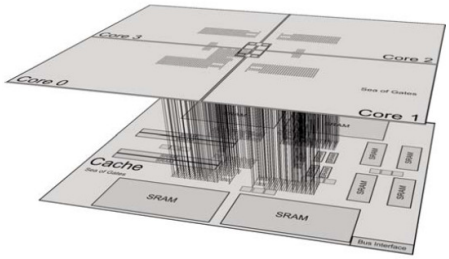
In [65], the design of 32 bit 3D adder (Kogge-Stone) and 32x32 3D multiplier (Wallace Tree) have been implemented using MITLL 180 nm 3D FDSOI technology to show the improvement of

arithmetic circuits in 3D architecture. The chip area is 1.3 mm x 1.3 mm die area running at 200 MHz based on post place and route timing estimation. The TSV size is 3 µm x 3 µm diameter and about 7 µm depth. The 3D adder showed up to 34% and 46% for speed improvement and power reduction while the 3D multiplier showed 14% and 7% of speed improvement and power reduction from simulation result as the fabricated chip is only used to prototype the idea and 3D design flow.

In [127], 3D SRAM is implemented using MITLL 180 nm FDSOI process showing 32% improvement of access time measured using delay-locked loop (DLL) owing to the reduced word-line wire in 3D architecture. The TSV size is 2.5 µm x 2.5 µm. The 3D SRAM has a 16 x 16 cell array in each tier using wordline split partitioning method. The design is tested at a range of 70 - 130 MHz to calculate the access time. The results of the measurement showed that 40 – 60 ps larger from the simulated result.

In [128], the design is implemented using three-tier MITLL 180 nm process in 6.3 µm x 6.4 µm die area. The design operates at 128 MHz achieving a throughput of 2 Gb/s with 430 mW power consumption. The 3D implementation shown significant improvement in terms of wire length, clock skew, area and buffer size over its corresponding 2D implementation. The 3D memory on memory architecture in 2.9 mm x 2.0 mm chip using Tezzaron two-tier technology with Global Foundries 130 nm technology has been implemented to demonstrate fast checkpointing and restore applications in 3D architecture [129]. Each SRAM tier has 1Mbit capacity built in 64 banks, each bank has 256 words and 64 bit wide. The chip can perform checkpointing/restart at 4k/cycles with 1 GHz speed.

In [83], a two-layers 3D multicore processor has been fabricated using Tezzaron technology as shown in Figure 34 based on 64 ARM Cortex-M3 processor using near-threshold computing (NTC) method to reduce power consumption in a 3D stacked system where the core operates at about 200 mV above its V_{th}. The design consists of 1591 vertical connections for the communication between a cluster of four cores to the adjacent caches. The cluster of four cores can operates at 10 MHz while the caches at 40 MHz. The fabricated chip is designed to be expandable to four layers of cores/caches with 2-3 layers of stacked DRAM. It can be achieved through taking two pairs of bonded cores/caches with face-to-face connection and stacked through back-to-back connection after the pairs is thinned to 12 µm on the caches side for TSV opening.



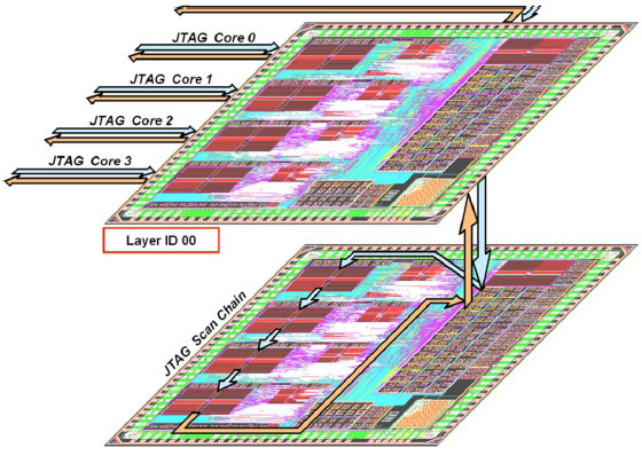
Logic Layer Dimensions	12.66 x 5mm
Technology	130nm
Metal Layers	5
Core Layer Devices	28.4 M
Cache Layer Devices	18.0 M
Core Layer Thickness	12um
F2F Connection Pitch	5 um
F2F Connections/Cluster	1591
Bus F2F Connections	2992
Total F2F Connections	28485
B2B Connection Pitch	5 um
Total B2B Connections	3024
DRAM Connection Pitch	25 um
DRAM Connections	3624

(a) (b)

Figure 34: Centip3De (a) architecture and (b) design summary

The most recent 3D architecture implementation is from [130] where homogeneous multiprocessor architecture (identical architecture for each die) has been fabricated in 2 tiers using UMC 90 nm technology demonstrating the modular 3D multiprocessor architecture which can be stacked in an arbitrarily number of layers in a single chip. The architecture as shown in Figure 35 is based on LEON3 commercial grade open source processors communicate using NoC arhictecture where each die has 4 processors that communicate using a single switch. The NoC is then connected to the serializer and deserializer (SerDes) macro block for inter-tier connection through in-house via-last Cu-TSVs to reduce silicon area occupied by the TSV. Although the TSV can be run at a higher frequency by increasing the clock frequency of serializer to compensate the bandwidth loss due to the serialization, it is not implemented in the fabricated chip because of silicon area limitation. The implementation of modular architecture together with SerDes macro block in each tier allows the architecture to be expanded in multiple layers to increase the design complexity without any additional modification.

Summary of the 3D chip implementation is shown in Table 12 for easy comparison in terms of different type of 3D technology and architecture used for each design. It is shown from the table that the widely used 3D technologies for the research chips are from Tezzaron and MIT Lincoln Lab using relatively older process technologies. Although it is better to use more advanced technologies such as 45 nm or 28 nm technology to demonstrate the real benefits of 3D technology because in advanced process technology wire delay has severe effect to the performance and power consumption, however, we can still achieve a significant performance improvement using old process technology for 3D architecture implementation by performing architectural specific manual optimization with the support of highly customized design tools at the expense of more design effort.



(a)

Process technology	90nm CMOS
Number of layers	2
Die size	4x4 mm
Core footprint	800x1650 μm
Number of cores	8
Total on-chip memory	320 KB
Max operative frequency	400 MHz
Vertical data bandwidth	3.2 Gbps
Number of I/O pads	120

(b)

Overall number of TSVs	120	(95∗)
Total Power TSVs	54	
Total Signal TSVs	66	(31∗)
• Total NoC TSVs	44	(22∗)
• Total JTAG TSVs	10	(5∗)
• Total Control Signal TSVs	12	(4∗)
TSV diameter	40	µm
TSV depth	50	µm
TSV capacitance	1	pF
TSV resistance	0.7	Ω

(c)

Figure 35: 3D modular multiprocessor (a) architecture (b) design summary (c) TSV parameters

Conclusion

This chapter reviewed 3D IC technology by initially discussing about the limitations of 2D architecture and CMOS scaling problems, and then provides motivation and potential benefits of 3D IC technology. Following that, various 3D technology aspects are discussed covering TSV technology, stacking methods, bonding orientations, manufacturing processes and 3D standards. Issues and challenges of 3D technology have been identified which require more research to be done before industry adoption become reality. Finally we presented some of the fabricated 3D test chips intended to provide a strong evidence of the 3D technology advantages in various architectures and application domains.

Table 12: 3D architecture implementations summary

No.	Work	Architecture	Purpose	Process technology	Number of tier
1	[124]	3D multicore	to demonstrate large memory bandwidth	130 nm	2 tiers
2	[84]	3D Mesh NoC with traffic generator	to demonstrate a working 3D NoC	180 nm	3 tiers
3	[59]	3D FFT processor	to demonstrate 3D benefits of speed improvement and area reduction	180 nm	3 tiers
4	[125]	3D SoC for H.264	to demonstrate 3D SoC architecture implementation	130 nm	5 tiers (2 tiers logic, 3 tiers DRAM)
5	[126]	3D NoC (single switch) with traffic generator	to demonstrate the feasibility of 3D NoC	130 nm	2 tiers
6	[65]	3D adder and 3D multiplier	to demonstrate arithmetic circuit improvement in 3D architecture	180 nm	3 tiers
7	[127]	3D SRAM	to demonstrate memory access time improvement in 3D architecture	180 nm	3 tiers
8	[128]	3D LDPC decoder	to demonstrate 3D architecture benefits (wirelength, power consumption, area)	180 nm	3 tiers
9	[129]	3D SRAM	to demonstrate fast checkpointing and restore application of hard disk drive	130 nm	2 tiers
10	[83]	3D multicore	to demonstrate low power 3D architecture	130 nm	2 tiers
11	[130]	3D multicore	to demonstrate modular multiprocessor architecture	90 nm	2 tiers

CHAPTER 4
3D DESIGN FLOW FOCUSING ON TIMING VERIFICATION

This chapter discusses a design methodology for 3D multiprocessor design concentrating on 3D timing verification. The proposed methodology leverages the benefits offer by Tezzaron two-tier face-to-face stacking technology using microbumps. Through this stacking method, it is possible to have 3D verification as early as at post synthesis stage due to the negligible delay of microbumps for the inter-tier communication. Having the advantage of 3D post synthesis verification allows us to explore the timing analysis several 3D architectural implementations and perform necessary modification to the design to maximize its performance compared with the 2D architecture counterparts by reducing costly iteration of timing exploration at place and route stage.

4.1 Related Works on 3D Design Flow

Design methodology to explore trade-off between timing, power and temperature in 3D integration based on MIT Lincoln Lab three-tier technology flow has been proposed by [131] as shown in Figure 36. Using two case studies of 8-point FFT and OpenRISC Platform System-on-Chip (ORPSOC) representing low power and high performance design respectively, they explore trade-off of leakage and dynamic power by varying the number of tiers from 1 to 10 tiers and thermal via density from 0% to 20%, where the thermal via is based on a standard cells design. K-Metis partitioning tool is used to partition the post-synthesis netlist to have minimum cuts for inter tiers connections. After floorplanning using Cadence SoC Encounter tool, thermal design is performed by inserting pre-determined number of thermal vias into the design where its location can be modified later to remove hotspot. Delay and power analysis is done after routing is completed followed by thermal simulation. Delay-temperature and leakage-temperature dependant behavior is analyzed to find the impact of temperature consideration when measuring timing and power of a 3D architecture.

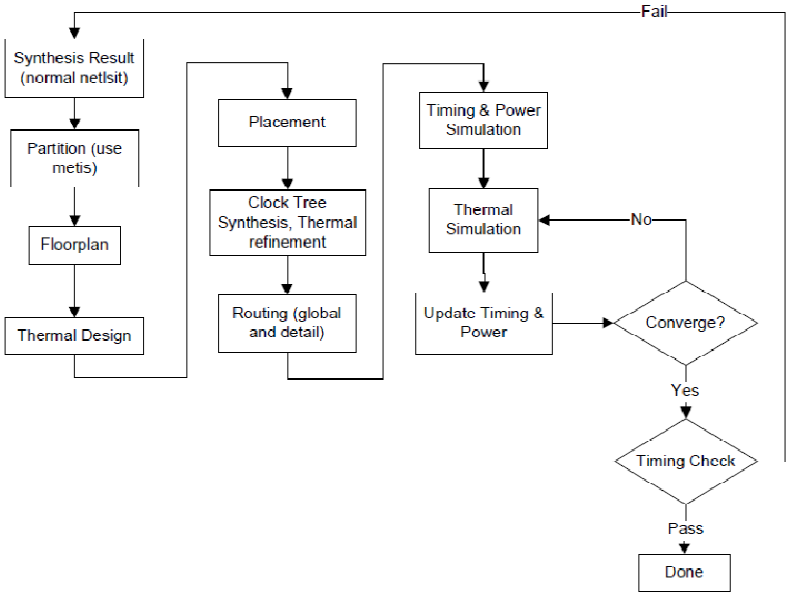


Figure 36: 3D design methodology for timing, power and temperature exploration [131]

Figure 37 shows a design flow for 3D ASIC design based on the standard supercell method presented by [132]. A new standard supercell layout scheme is proposed following the issues related to the TSV such as its large size and the fact that specific TSV-placement strategy is needed to reduce the routing density to be able to target for interconnect-heavy VLSI circuits implementation. The standard supercells are basically the layout macros with the same height and varying widths which is placed in rows with spaces are reserved between the macros for 3D via and buffers to enable easier handling of 2D as well as 3D wires during the physical design implementation. Concerning the 3D design flow, 3D specific design tools have been introduced based on the proposed standard supercell layout scheme including 3D-placer and 3D-via assignment tools co-design with several other tools that are based on the existing commercial 2D EDA tools such as 3D buffer insertion, 3D clock distribution and 3D LVS.

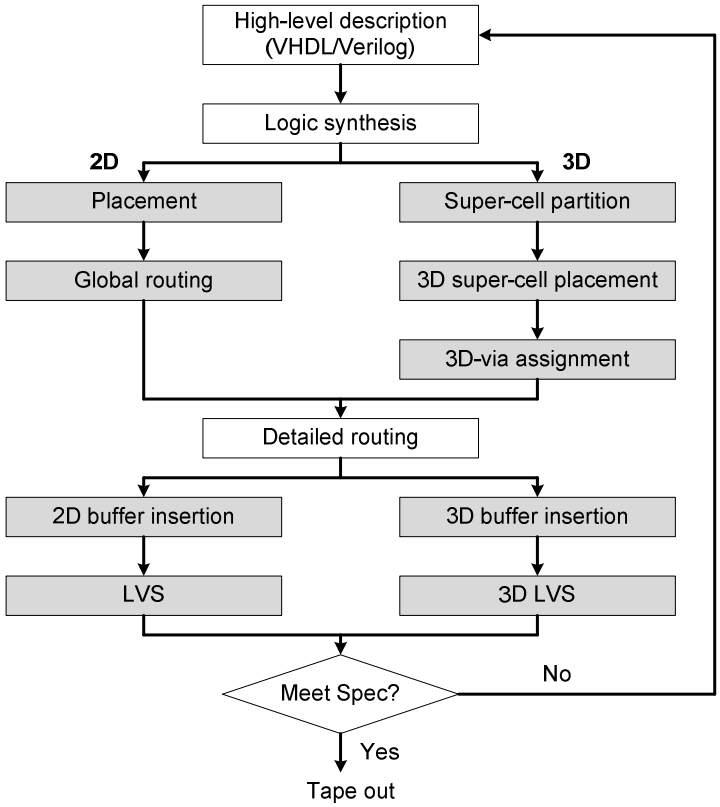


Figure 37: 3D ASIC design flow based on standard supercell layout [132]

A CAD flow for via-last 3D integration using face-to-back stacking is proposed by [133] shown in Figure 38. Vertical interconnect is inserted into the synthesized netlist. During the floorplanning stage, manual partitioning is done by firstly considering the TSV power and ground connections. Next, TSV-aware mixed-sized placement is performed followed by multi-tier CTS and routing. DRC and LVS are performed after merging the routed GSII files. The methodology is evaluated using several ISCAS89 circuit benchmarks implemented in two to four tiers stacking based on TSMC 90 nm standard library technology.

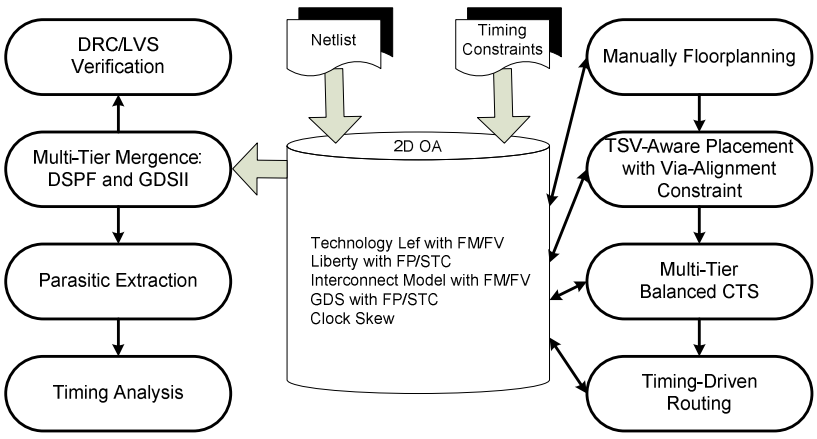


Figure 38: CAD flow for via-last face-to-back 3D integration [133]

Another design flow as shown in Figure 39 is proposed by [134] for designing FFT architecture. The design is partitioned initially using K-Metis tools into three-tier architecture. Placement stage is done iteratively where modification is done manually after every placement to get a better floorplan by firstly fixing the FIFO memory location and then the TSV in the middle tier and copying their locations to the other two tiers. Next, the rest of the standard cells is placed and routed independently for each tier. Finally, the netlist and SPEF files of the three tiers are extracted for power and timing analysis.

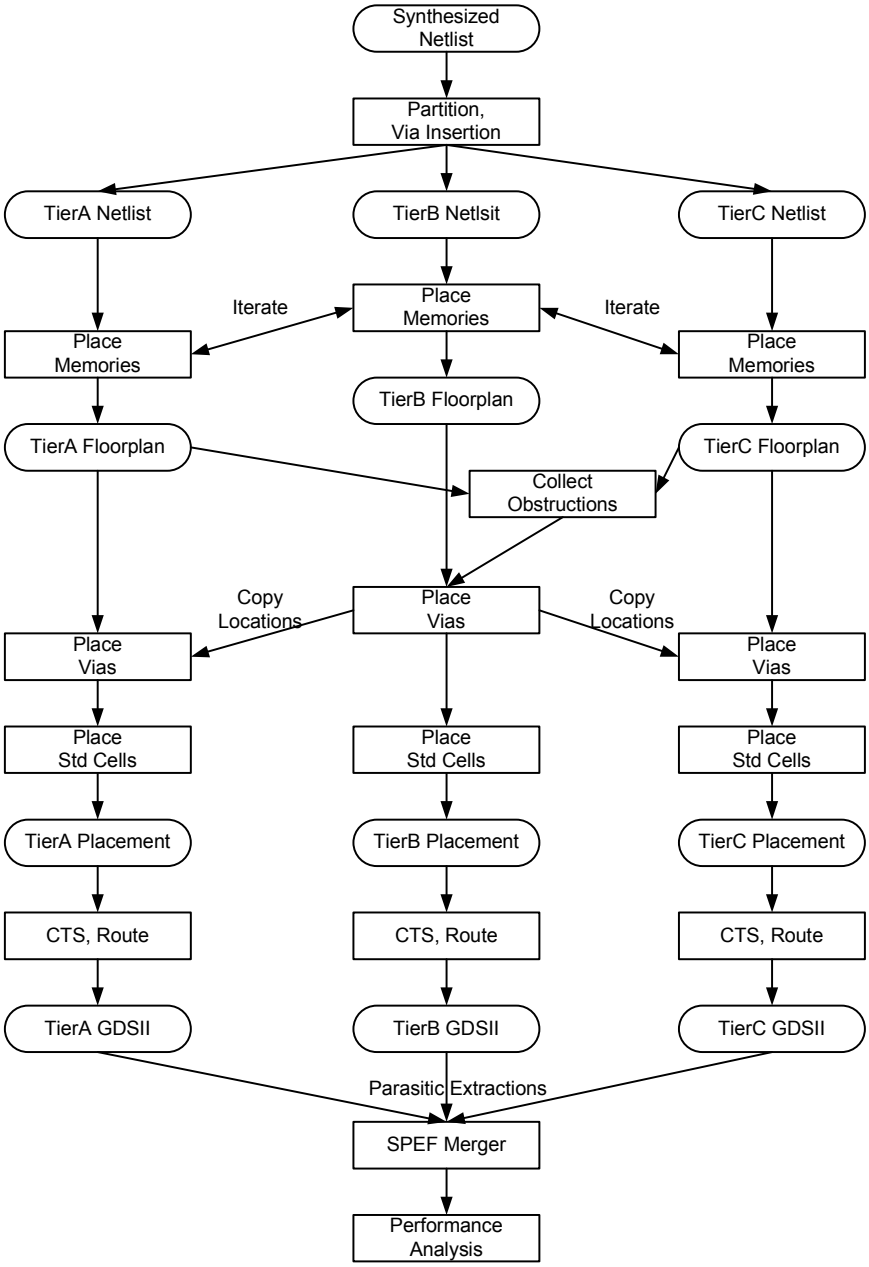


Figure 39: 3D design flow for three-tier FFT architecture using MIT Lincoln Lab technology [134]

An automated design flow for 3D microarchitecture performance evaluation flow is reported in [135] to have qualitative result of performance and thermal properties to aid 3D architecture design. The inputs of the flow are several microarchitecture properties such as operating frequency and estimated power density of blocks. The automated floorplanner can performed 2D and 3D floorplanning blocks which can be configured to optimize area, performance and temperature. Thermal via insertion and global routing is then performed to further optimize the thermal and routing profile. The generated netlist is then fed into cycle accurate simulator that is used to evaluate latency, power and thermal of the design to validate the result of the flow. The proposed design flow shown in Figure 40 is validated by using an out-of-order superscalar processor case study.

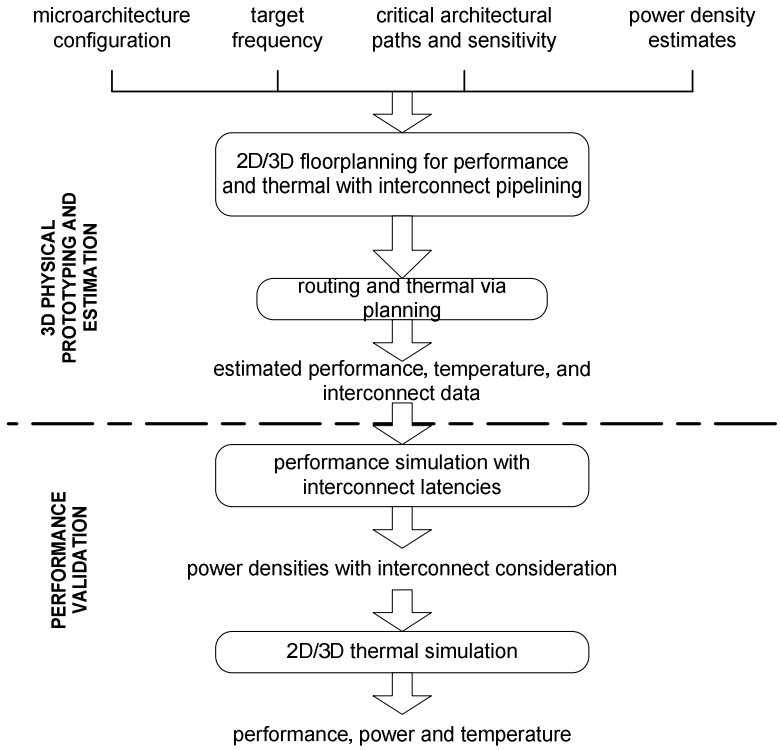


Figure 40: Automatic design for 3D microarchitecture performance evaluation [135]

A design flow to implement Synthetic Aperture Radar (SAR) processor has been presented in [59] using three-tier MIT LL 3D technology. The proposed design flow shown in Figure 41 is customized with the manual floorplanning of SRAM and ROM for the FFT architecture which is the main elements in the SAR processor. The three tiers of the design are synthesized separatedly after the floorplaning is completed. During place and route stage, manual floorplanning of the entire architecture is done by manipulating the DEF files in the SoC Encounter to determine the TSVs locations. The TSV for inter-tier connections is manually inserted by hand in the Virtusos layout editor due to the small number of TSV requires in the design (24 TSVs). Using this design flow,

they exploit performance improvement using 3D technology by predetermined the 3D floorplanning before continuing the front-end and back-end design flow whereas no timing optimization has been taken into consideration during the flow.

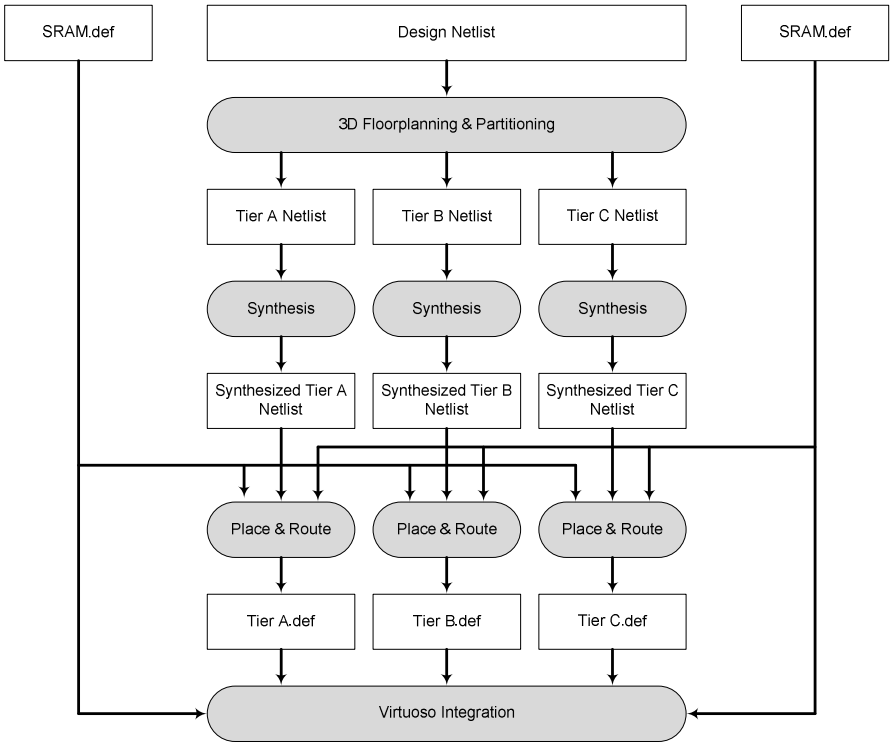


Figure 41: Design flow for 3D SAR processor [59]

The most recent 3D IC design flow is proposed by [69] to design hybrid process of 3D architecture shown in Figure 42. The design flow is based on heterogeneous 3D integration implementation where two dies consists of core processor that is fabricated using high end process technology and IO circuit (with other low speed circuit such as memory blocks) that is fabricated using low end process technology is stacked face-to-face using microbumps to reduce manufacturing cost while preserving optimal process technology for each logic function.

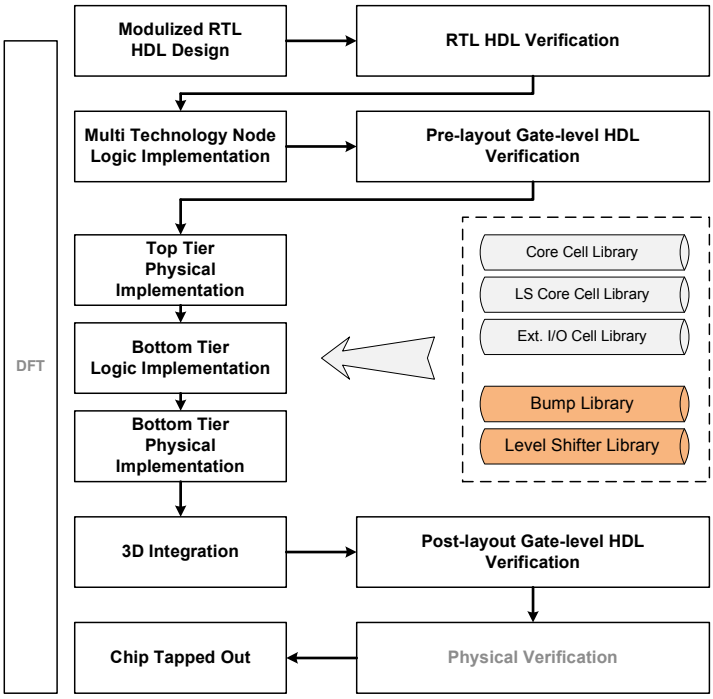


Figure 42: Design flow for 3D hybrid process architecture [69]

4.2 Proposed 3D Design Flow with Timing Verification

4.2.1 3D Physical Design Implementation Flow

The proposed 3D design flow in our work is shown in Figure 43 leveraging the small delay of inter-tier connections due to the microbumps structure. This flow is made possible with the 3D technology such as offered by Tezzaron technology which uses two tiers with face-to-face stacking using mirobumps for inter-tier connections. This generic 3D design flow can be reused for any 3D architecture targeting the Tezzaron 3D technology unlike some of reported design flows that are quite complicate to be implemented and tailored to a specific design and also require many in-house scripts [136]. Additionally, compared with previous design flows, we perform 3D verification in the back-end and front-end. These microbumps structures have negligible delay for the inter-tier connection and thus we can perform 3D timing analysis at post-synthesis stage without any inaccurate delay estimation of inter-tier connection. Therefore we can have an early timing performance estimation of the 3D design after synthesis stage and save time because we can have architectural modification to satisfy performance specification before proceeding to the place and route stage as it takes quite long run time particularly for a relatively large design with very high number of mirobumps as well as to fix other DRC violations.

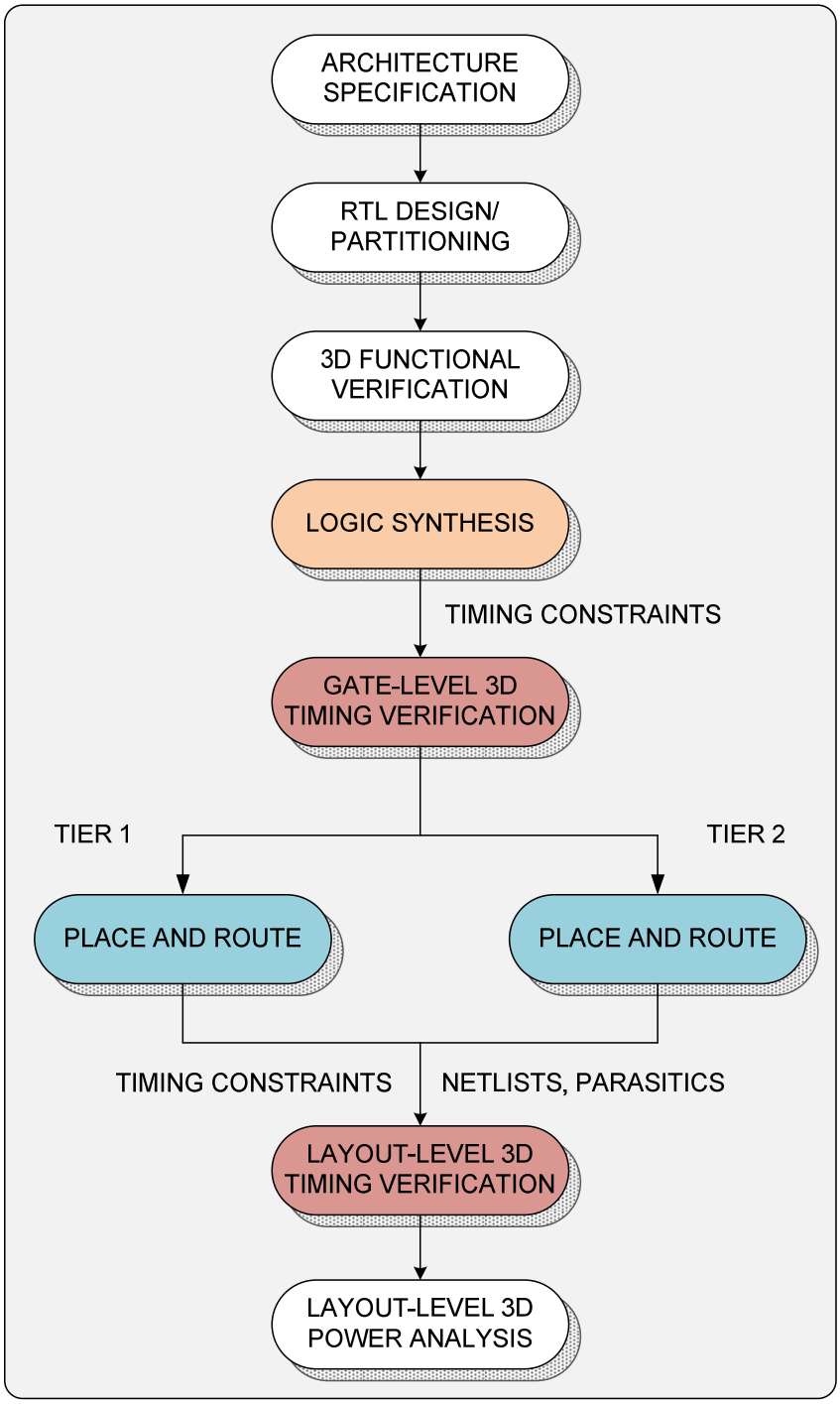


Figure 43: 3D design flow focusing on timing verification

4.2.2 Front-end Design Flow

The design is initially partitioned into two blocks corresponding to two tiers at RTL level. Following that, front-end design flow is carried out using Synopsys Design Compiler with timing budgeting flow shown in Figure 44. Timing budgeting flow is a method of distributing timing constraints among logic blocks (in our case the partition blocks) so that each block can be separately implemented and optimized by their own back end flow. First, the top level 3D design containing two partitioned blocks is analyzed and elaborated before 3D timing constraints is applied. Next, timing budgeting command is executed to generate timing constraints for each partition where it will be used to compile and to generate netlist of each partition. Once the compilation of partitioned blocks is completed, compilation of top level 3D design is then performed and timing is analyzed. In case of timing violations exist at each partition block or at top level 3D design, the 3D timing constraints is modified by relaxing the clock frequency and the bugeting flow is repeated. The timing budgeting flow is only feasible to be employed due to the small/negilible delay of microbumps structure for vertical connection whereas not 3D technology based on TSV which will produce less accurate budegeted timing constraints as synthesis tool does aware about inter-block delay model for the TSV. The accuracy of budgeted constraints could be even worse when using multiple TSVs connection per net in the 3D architecture.

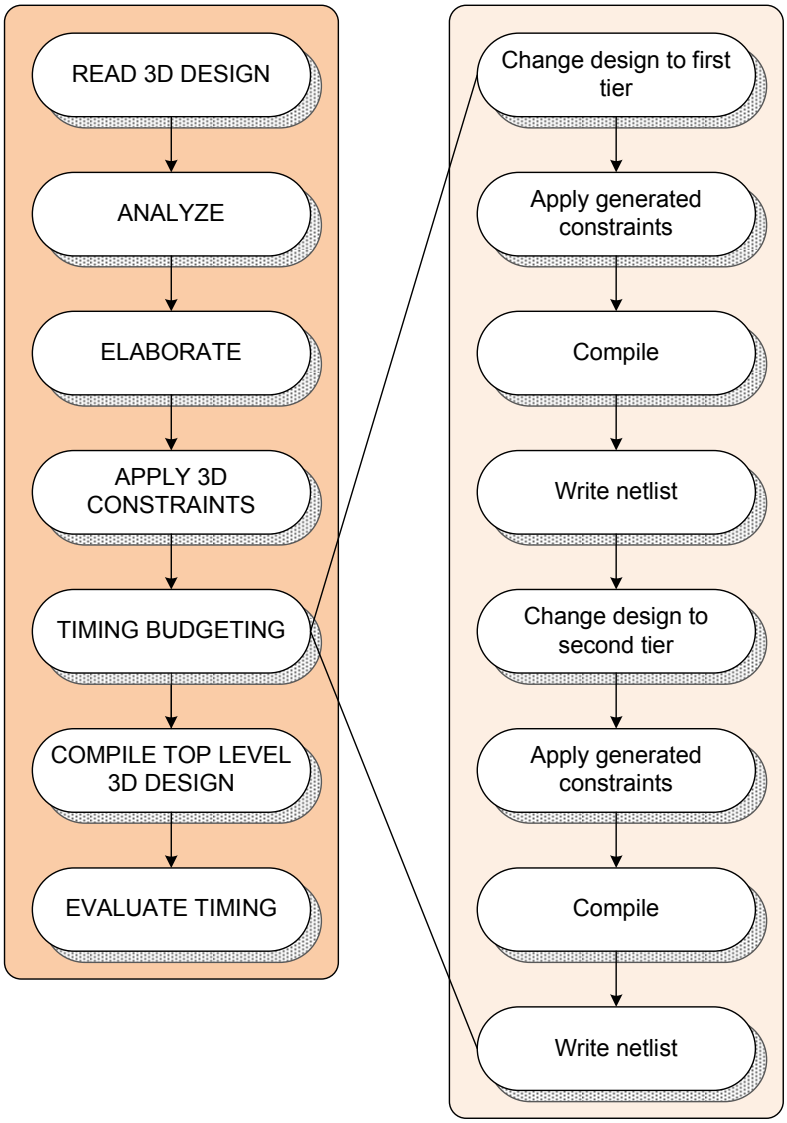


Figure 44: Front-end design flow with timing budgeting flow

4.2.3 Back-end Design Flow

From the generated gate-level netlists and timing constraints of each partition, place and route step is performed. Figure 45 shows the detail place and route flow where it is like normal 2D design flow except the additional microbumps insertion for the inter-tier connections including steps for bumps assignments for inter-tier connections. The location of the microbumps is determined manually by looking at the logic block it will be connected in order not to have long horizontal length before reaching the vertical interconnection. Once the location is fixed, arrays of bumps are created and then the numbering of the bumps is modified so that it is easier to assign to the signals. Finally, physical pins are created under each microbump to be able to route by the router in SoC Encounter (NanoRoute). During floorplanning stage, we need to capture the location of microbumps for inter tier connection such that we could connect the same signals to the same microbumps on the other tier with mirroring. The microbumps for power delivery structure must also honors their location between both tiers. During the power planning step, a vast array of microbumps is formed for a given power ground wires such that enough current can be supplied to the other tiers to ensure correct operation. The location of microbumps for the power ground connection is aligned between both tiers such that it will be overlapped on top of each other once the stacking is done. It is followed by the conventional 2D design flow which is placement, clock tree synthesis and routing with optimization is guided by the generated timing constraints from front-end flow.

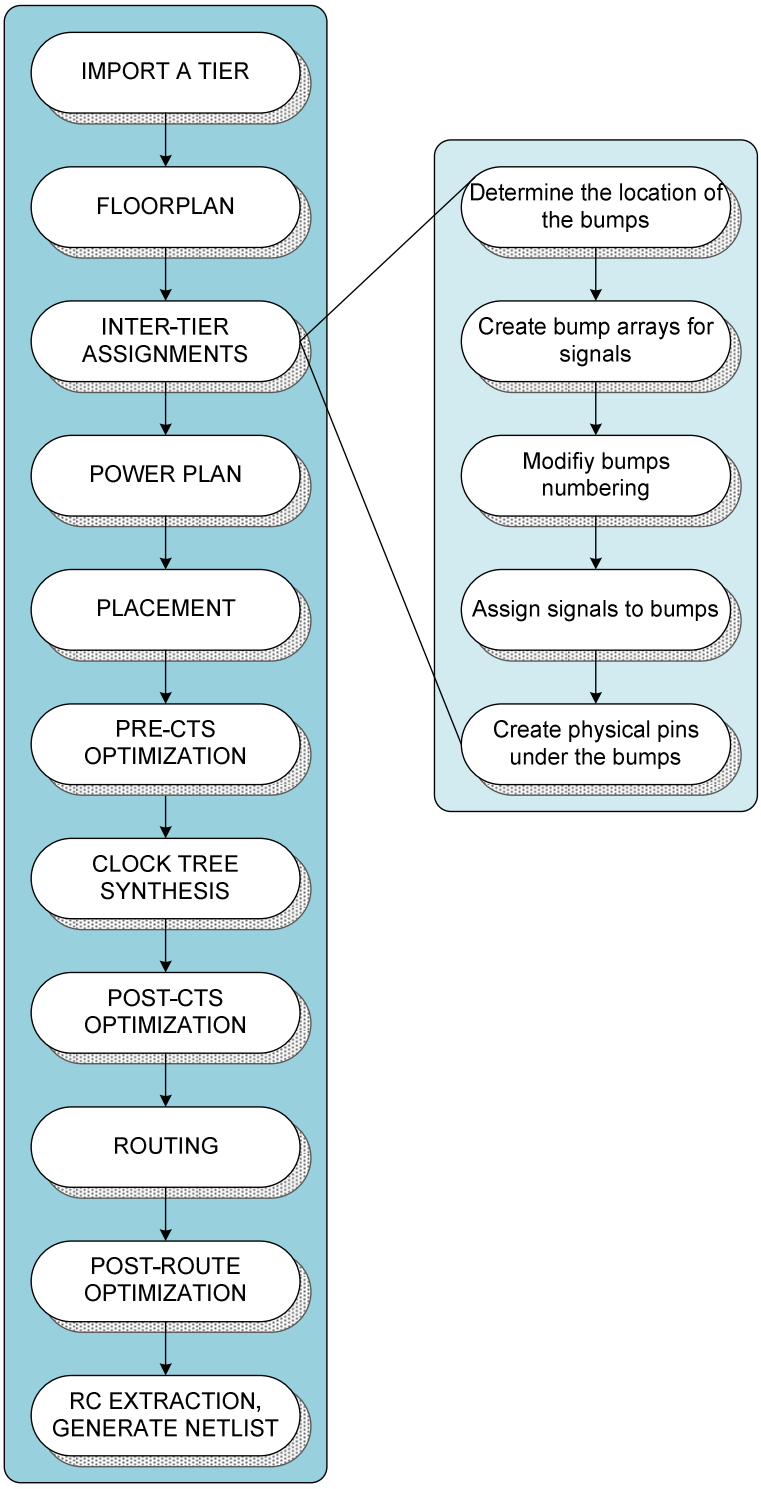


Figure 45: Back-end design flow with inter-tier signal assignments

4.2.4 3D Timing Analysis Design Flow

Figure 46 shows 3D timing analysis for gate-level and layout-level correspond to the general 3D design flow shown in Figure 43 using Synopsys PrimeTime tool and PrimeTime PX for power analysis. The 3D timing constraints is used for gate-level 3D timing verification while for layout-level 3D timing analysis and power analysis, additional RC parasitic files generated from SoC Encounter is used by reading it into PrimeTime incrementally one tier after another. The parasitic value of microbumps is set to zero when issuing command to read parasitics file. 3D timing paths can be mainly classified into several groups as follows:

1. Intra-tier paths: timing path from a flip-flop located in a layer, through the combinational logic gates and ends a flip-flop located in the same layer. These timing paths will be automatically optimized by the 2D implementation tools as in conventional physical design flow using the same timing constraints as in synthesis flow.
2. Inter-tier paths: timing path from a flip-flop located in a layer, through the combinational logic gates, microbumps or TSVs and ends at a flip-flop located in another layer. 2D place and route tools do not see these paths and thus do not able to optimize it. Although constraints have been budgeted for each tier during the synthesis flow, the timing optimization during place and route flow is not aware about the nets connected at the other tier making top level 3D timing closure difficult.

Within the aforementioned timing path groups, it can be further classified into intra-block (inside a block within the top level implementation) paths and inter-block (between different blocks). Referring to the tile-based design methodology, the paths can be within-tile paths and tile-to-tile paths as discussed in [137]. Unless these paths are in the separate tier connected vertically, thus timing closure is very likely to be achieved using convential mature place and route tools.

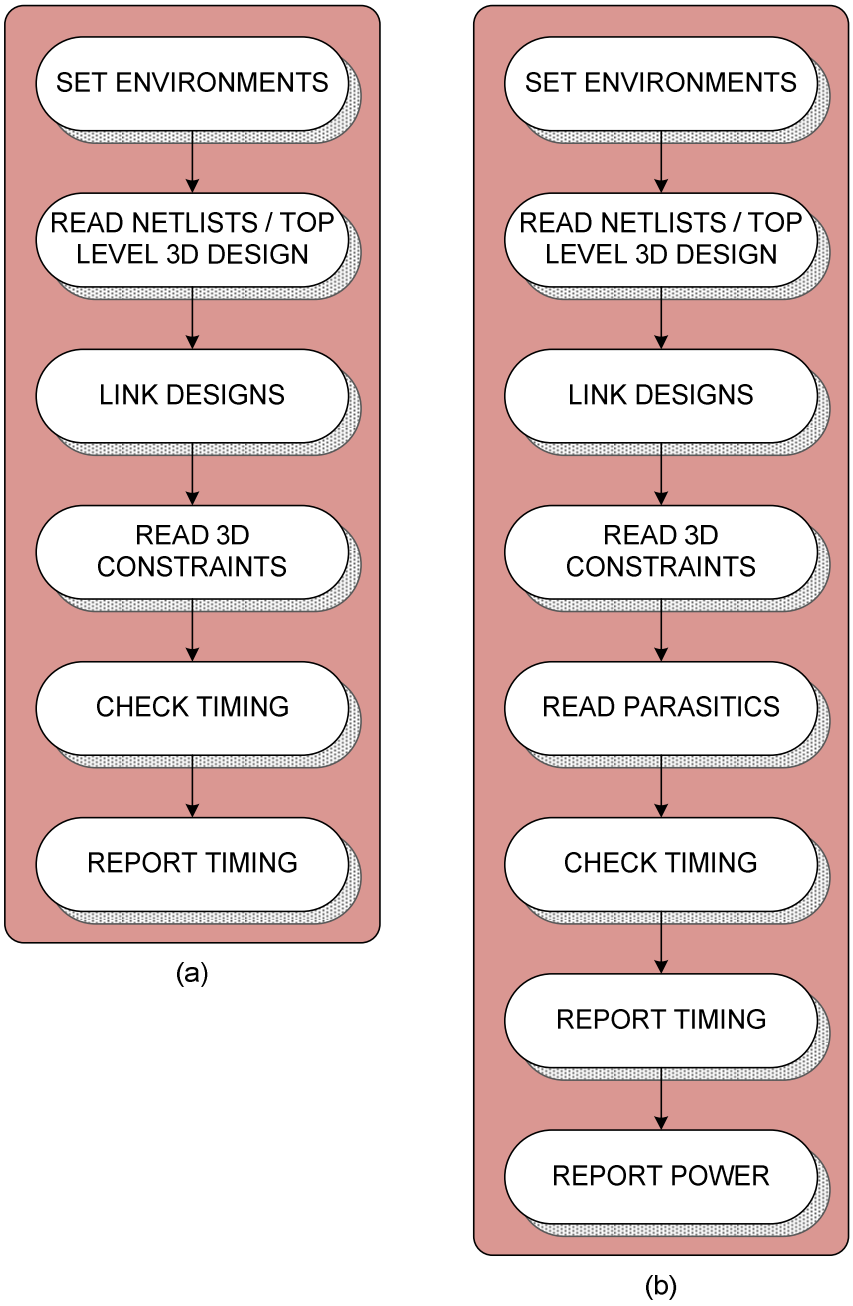


Figure 46: 3D timing analysis (a) gate-level (b) layout-level with power analysis

4.2.5 Limitation of the Flow

There are several limitations imposed by this flow. First, this flow can only be used for two-tier face-to-face integration such as in Tezzaron 3D technology. This is due to the timing budgeting method we use for allocating the timing between the two tiers. When using TSV for stacking, 3D timing verification may not be accurate. Since mirobumps have very small delay, thus the Synopsys Design Compiler can allocate the timing between the two blocks accurately. If TSV is used for the stacking, the budgeting process may not be accurate because the TSV delay must to be considered which will depend on the target TSV properties such as diameter, height, pitch, material (TSV, insulator). This TSV delay is not known during the synthesis process. Second, this flow requires the 3D design to be manually partitioning into two blocks at RTL level prior to run the synthesis step to be able to apply the timing budgeting process. This requirement prevents designers from of automatic 3D partitioning tools capable of increasing 3D architecture performance.

Conclusion

In this chapter, we have discussed the state of the art of design methodologies for 3D integration considering their advantages and disadvantages. Then we explained our proposed novel 3D design methodology concentrating at 3D timing verification methodology. The design flow is specific to the Tezzaron two-tier 3D technology which is based on microbumps structure for inter-tier connection but can be applied for any other 3D designs to be implemented on the same technology. Leveraging from its small structure and therefore its negligible delay, the proposed design flow is able to identify 3D timing path in early design cycle and thus making make modification to be able to reduce costly iteration at place and route step.

CHAPTER 5
EXPLORATION OF 3D NOC ARCHITECTURES THROUGH PHYSICAL DESIGN IMPLEMENTATION

The need for physical design implementation for evaluating 3D NoC performance is important to provide more accurate analysis on how much benefits could 3D integration bring to the NoC architecture than performance analysis through simulation methodology as reported by several works previously. In this chapter, we perform an exploration of novel 3D NoC architectures through physical design implementation based on two tiers Tezzaron 3D technology. The 3D NoC partitioning is done by dividing the NoC's datapath component into two blocks to be placed in the two tiers. Two 3D NoC architectures namely 3D Stacked Mesh NoC and 3D Stacked Hexagonal NoC developed based on this partitioning strategy are analyzed by comparing its performance with 2D Mesh NoC and its straight forward 3D extension, 3D Mesh NoC. In order to measure the impact of wire delay on its performance, two standard cell libraries (130 nm and 45 nm) representing old and advanced technologies have been used for the performance analysis. Results from physical implementations show that in advanced technologies such 45 nm and below, the performance of 3D NoC with datapath partitioning method has better performance compared with traditional 2D/3D NoC architecture.

Our motivation is that we want to explore different partitioning strategies from the previous reported works for 3D NoC architecture and then evaluate its performance accurately based on layout-level routed netlist. The contributions in this chapter are as follows:

- Propose a new partitioning strategy which is based on partitioning router's datapath component into two blocks to be placed in two tiers based on Tezzaron 3D technology. The performance of the NoC architectures based on this partitioning method, namely 3D Stacked Mesh NoC and 3D Stacked Hexagonal NoC are evaluated and analyzed by comparing with 2D Mesh NoC and 3D Mesh NoC.
- Study the wire delay impact on the 3D NoC architectures by performing physical implementation using two standard cell libraries, 130 nm and 45 nm representing old and advanced process technologies. Gate delay is dominance is old technology while for advanced technology, wire delay have significant impact on the performance.

- Provide detailed 3D performance analysis for 3D NoC architectures based on the result from layout-level routed netlist. We highlight some of the key parameters contributing to the performance of 3D architecture based on the physical implementation results and also discuss the performance impact of using 2D EDA tool for 3D implementation.

5.1 Related Works

Many issues in 2D NoC architecture and design have been studied over the past several years covering various aspects such as design flow, implementation evaluation and design space exploration [138] [29] [139] [140] [141] [142]. However, research in 3D NoC is still in its infancy and many issues remain unexplored especially in real design and implementation. Design space exploration of 3D NoC topologies through cycle accurate simulation has been performed previously showing the benefits of 3D design in terms of throughput, latency and energy dissipation for mesh-based and tree-based NoC architectures [143]. 3D NoC architectural study using a combination of heterogeneous floorplans for logic blocks and homogeneous floorplan for mesh NoC through software simulation analysis is presented in [144] showing the benefits of the proposed architecture over conventional 2D design. The 3D Stacked NoC architecture is proposed in [145] where routers is stacked into multiple layers to optimize power and reduce overall area instead of implementing 3D routers with additional ports for up and down links. Performance analysis through cycle-accurate simulation has proved the benefit of the proposed architectures compared with normal 3D router. In [146], zero load latency and power consumption analytical model of various 3D NoC topologies are evaluated proving the advantages of combining 3D IC with 3D NoC architecture. Their work showed that by reducing the physical area of the tiles it could gives speed improvement because of the inter-router wire length reduction while power consumption is reduced due to the reduction of the number of hops for 3D mesh NoC architecture. We based upon their paper to investigate further the results by doing analysis from physical design implementation results.

As for the architectural study, a design of asynchronous 3D router has been proposed to optimize TSV utilization using serializing method providing higher speed for inter-die packet transfers as well as increasing the throughput while added power and area overhead for serialization and deserialization logic [147] [148]. Another work proposed a novel 3D router architecture by decomposing the router into different dimensions to provide better performance over other 3D NoC architectures [149]. 3D architectures of crossbar and multistage interconnection network are presented in [150] showing performance improvement over its 2D structure. TSV sharing over time division multiplexing technique for vertical interconnection ports between routers is proposed by

[151] to improve TSV utilization which eventually reduce the number of TSVs specifically for symmetric 3D mesh NoC. Differs from the previous reported works, this work focuses on partitioning of 3D NoC architectures and evaluating its performance through layout-level netlist for more accurate analysis of wirelength, timing, area and power consumption.

Several experiments have been conducted investigating the performance of 3D architectures based on the results from physical design implementations. Work in [152] has studied different partitioning styles for implementing 3D multicore architecture namely core level, block level and gate level showing that TSV capacitance, EDA tools and timing optimization methods have strong impact on the performance of the final 3D architecture. In [153], they showed that 3D architectures could lose or reduce its benefits due to the tools inability to perform 3D-aware optimization. On the other hand, larger circuits tend to gain more improvement from 3D architecture over its 2D counterpart for advanced technology such as 45 nm node. In [154], the study of different 3D placement methods on the performance of three 3D architectures showed that true-3D placement method produces the highest performance improvement over other methods at old process technology (130 nm) indicating the importance of 3D-aware tools to obtain maximum benefits of 3D integration. However, no previous work has been presented with detailed performance evaluation on various physical design metrics (wirelength, timing, impact of wire length) of 3D NoC architecture in particular 3D Mesh-based NoC architecture.

In terms of partitioning technique for 3D architecture, different architectures pose different partitioning techniques to have optimal benefit out of stacking structure. Memory architecture which tends to have regular structure consists of regular array of cells that can be easily partition into several tiers by dividing either its bitlines or wordlines and have been demonstrated to have up to 31% performance improvement in two tiers over 2D design [155]. For logic design that has irregular structure, automatic partitioning tool such as hMetis was used showing 28% performance improvement in three tiers design [134]. Another work partitioning a large design into two tiers by manually determining the location of functional unit blocks in both tiers [125]. However, none of the previous work demonstrated significant performance improvement especially in terms of speed (ideally $\sqrt{3}$ improvement for three tiers design and $\sqrt{2}$ improvement for two tiers design) because the partitioning method did not consider 3D critical paths as it should be to have maximum performance. Compared with previous works, we do not use automatic partitioning tool in this study. Rather, we use manual partitioning method to partition the NoC architecture into 3D NoC architectures.

Generally, from designers' point of view, 3D technology can be viewed from technological and architectural perspective. Choices such as inter-tier connections using microbumps or TSV and number of tiers to be stacked refer to the technological constraints while partitioning method and TSV placement technique refer to architectural constraints. Tezzaron [99], MIT Lincoln Lab [96], Ziptronix [95] and IMEC [94] are some of the technology providers offering various types of 3D technology such as bonding technique, stacking orientation and number of tiers for commercialize as well as research purposes. As we have fixed the technological constraint in this study which is using microbumps for inter-tier connections, we conduct experiments by considering architectural constraint of 3D technology by analyzing the impact of physical area of each functional block (tile area) to the delay.

5.2 Standard Cell Libraries

This 3D integration technology is based on Tezzaron [156] that uses TSV for peripheral IOs. Table 13 shows the detail parameters for this technology use in this design. The two-tier 3D stacking method is based on wafer-to-wafer bonding, face-to-face method with via-first approach. Inter-die connection is achieved through microbumps structure where it provides high interconnection density up to 40,000 per mm^2 without interfering to FEOL (front-end-of-line) device or any routing layers. Furthermore, as its physical structure is small enough that the delay can be negligible, 3D verification methodology at every stage of physical design flow can be performed to estimate the design performance at early stage of the design and then do modification according to the specifications. It is also possible to implement four tiers design by stacking through back-to-back using TSV of the two face-to-face stacking in order to have higher design complexity.

Table 13: Physical design parameters using 130 nm standard library

Parameters	Details
Standard cells	Global Foundries 130 nm
Supply voltage	1.5 V core, 3.3 V I/O
Metal layers	6 layers (metal 6 reserved for bonding)
3D technology	2 tiers face to face bonding, wafer to wafer
TSV technology	Peripheral I/O, outside interface
TSV dimension	1.2 µm diameter, 2.5 µm pitch, 6 µm depth
Microbump dimension	3.4 µm diameter, 5 µm pitch

In order to analyze performance of 3D NoC architectures in advanced technology, we have chosen 45 nm standard cells from ST Microelectronic [157]. We use similar 3D structure for inter-tier connections using microbumps as in Tezzaron technology but we replace the 130 nm technology of Global Foundries with 45 nm ST Microelectronic standard cells summarized in Table 14. The 45 nm technology use in this study has seven metal layers where metal seven is used for bonding and the routing is done until metal six.

Table 14: Physical design parameters using 45 nm standard library

Parameters	Details
Standard cells	ST Microelectronic 45 nm
Metal layers	7 layers (metal 7 for bonding)
Voltage supply	1.1 V core
3D technology	2 tiers face to face bonding, wafer to wafer
Microbump dimension	3.4 µm diameter, 5 µm pitch

Previous works have shown that we will benefit more when increasing the number of 3D layers, however the most significant improvement is achieved when going from single layer (2D architecture) to two layers [60]. If we partition the design further to for example three or four layers, the performance improvement from adding second layer to third layer is smaller than from single layer to two layers. Similar trend happens when we move from third layer to fouth layer where the performance improvement become less than moving from second to third layer. However, this could be different if true 3D tools are used for physical design which can optimize the 3D paths in all tiers. With that in mind and considering optimal benefit of microbump that does not block routing wires for inter-tier connections, we have chosen two tiers for the 3D NoC architecture exploration in this study.

5.3 Baseline NoC Architecture

Before we explain the 3D stacked NoC architectures, we first describe the router, NIU architecture and the baseline 3D Mesh NoC in this section.

5.3.1 Router and NIU Architecture

The router and network interfaces are designed using VHDL. The functional simulation of the NoC is performed using Modelsim tool while logic synthesis and place and route is carried out using

Synopsys Design Compiler and SoC Encounter respectively as explained in the previous section. For all 3D NoC architectures discussed in this work, we set the initial target utilization to 60% - 65% for the NIU and the router during place and route stage. Placement constraints are used to do the floorplan of each tile for all designs rather than block level implementation to save time.

The network interface architecture as shown in Figure 47 is connected to the processor through two First-In-First-Out (FIFO) ports. Based on data address and number of words sent by the processor through one of the FIFO port, the network interface will access the processors data memory to process data blocks through DMA. Each network interface unit is connected to the processor through 2 FSL ports (FIFO) of the Openfire processor; one is master FSL for writing data to be transferred through the NoC and the other one is slave FSL for reading synchronization flags sent by other processors. The synchronization FIFO has 16 words (one word per processor) with 5 bits data width each. There is one 11 bits counter in the network interface unit for measuring packets travel timing. The timing information is included in the head flit attached to the packets when entering the network and is processed when the packets arrive at the destination network interface.

The input buffered-based 3D router architecture shown in Figure 48 comprises four neighboring ports, one vertical port for connection to another tier and one local port to the processor through network interface unit. Each input/output port has 35 bits data flits and 2 bits control signals for packet transfer between routers. Handshake protocol is used for router to router communication and router to network interface communication. Each input port has one buffer built using 16 words FIFO based dual port Synchronous Random Access Memory (SRAM) architecture to support a maximum of 16 data blocks transfer. As XY routing is deadlock free and we do not implement priority packets transfer, virtual channel implementation is not necessary. We use round robin arbitration for output port selection when there is more than one input requesting the same output route. Wormhole switching is used for packet transfer in the NoC because it does not require large buffer and has lower latency. For the routing, deterministic coordinate-based routing is implemented using XYZ coordinate where each packet will travel first in the X direction followed by Y direction and finally through Z direction (vertical) to the other die. Router vertical connections are implemented physically using microbumps at top most metal layer.

The packet format shown in Figure 49 consists of several flits depending on the number requested by the processor with up to 16 data flits at maximum. The head flit includes source and target router address, number of words to be transferred and synchronization bits. The first body flit carries memory address to write the data at the destination memory. In addition timing information

generated by a counter is also attached to this flit for evaluating packet travel time.

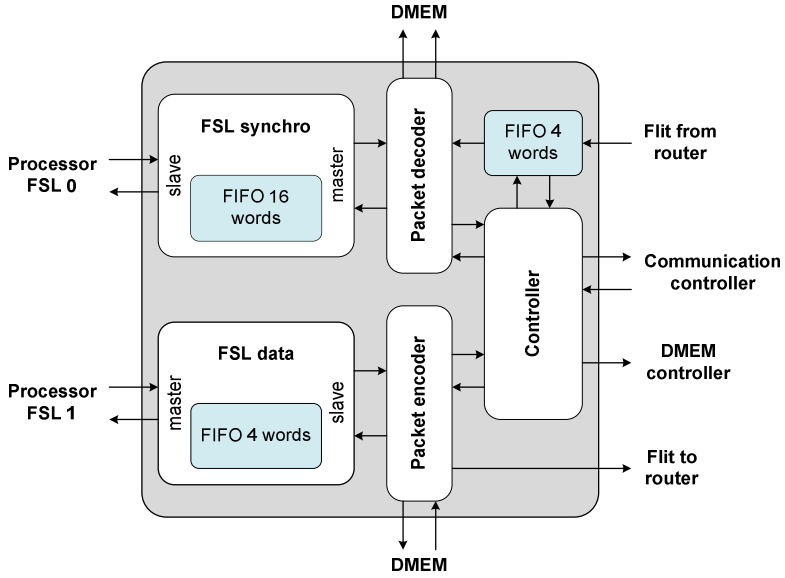


Figure 47: NIU architecture

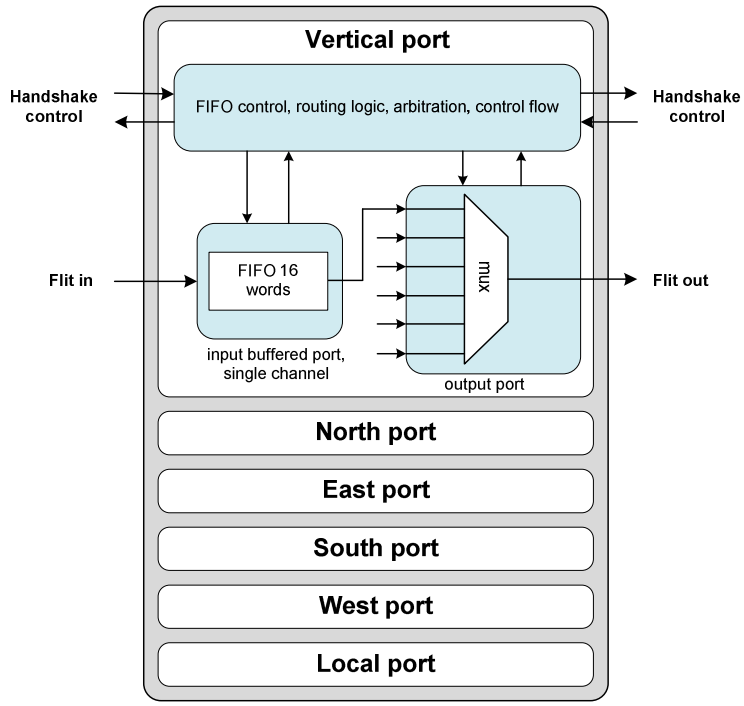


Figure 48: 3D Router architecture

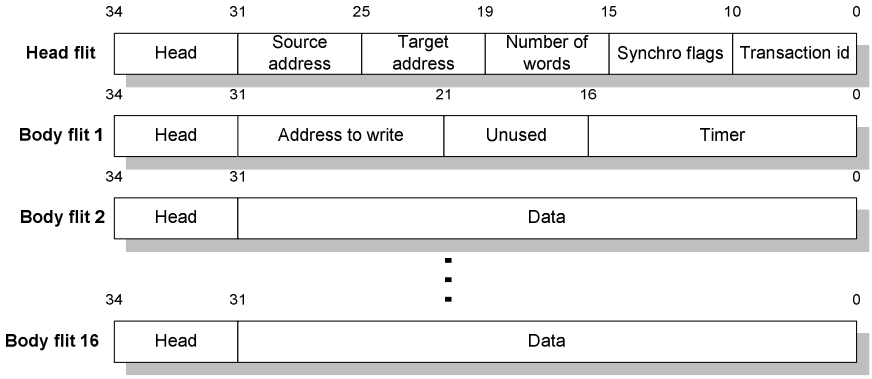


Figure 49: Packet format of the NoC

5.3.2 Baseline 3D Mesh NoC

In this architecture, the 3D NoC is implemented on two tiers where each tier has identical blocks as shown in Figure 50, Figure 51 and Figure 52. This is the straight forward extension of 2D Mesh NoC architecture where we just take a copy of a tile (a router and a NIU) and put it on top of another tile. Compared with the area of 2D Stacked Mesh NoC, this architecture has slightly more area due to the additional ports for vertical connections. This 4x2x2 mesh NoC architecture is based on 3D router architecture that has vertical links for inter-tier connections between routers. These physical vertical links shown in red color is based on the logical vertical links in each 3D router. This simple 3D NoC reduces chip area by half over its 2D architecture in 130 nm technology but is slightly increased for 45 nm technology (shown in Figure 61 and Figure 63). It provides latency improvement through reducing its network diameter (reducing number of hops through vertical links) from six to five hops. The problem of this architecture is that the inter-router physical links between horizontal and vertical links is unequal and thus this topology is not an optimal solution because the benefits is only from the shorter vertical inter-router links while its horizontal links remains the same as in the 2D topology.

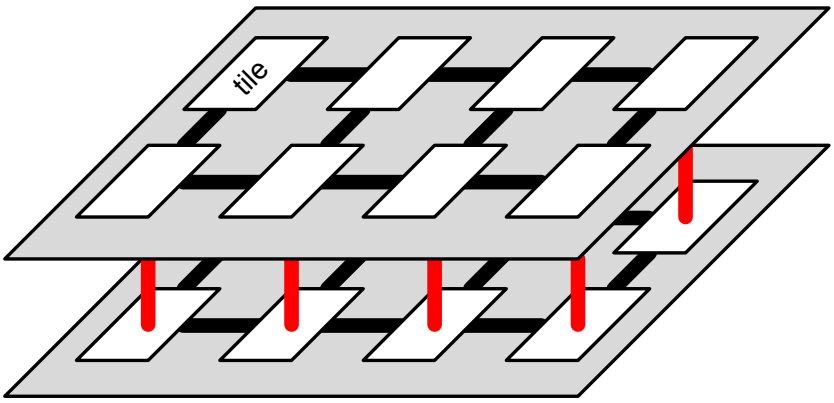


Figure 50: Block diagram of 3D Mesh NoC

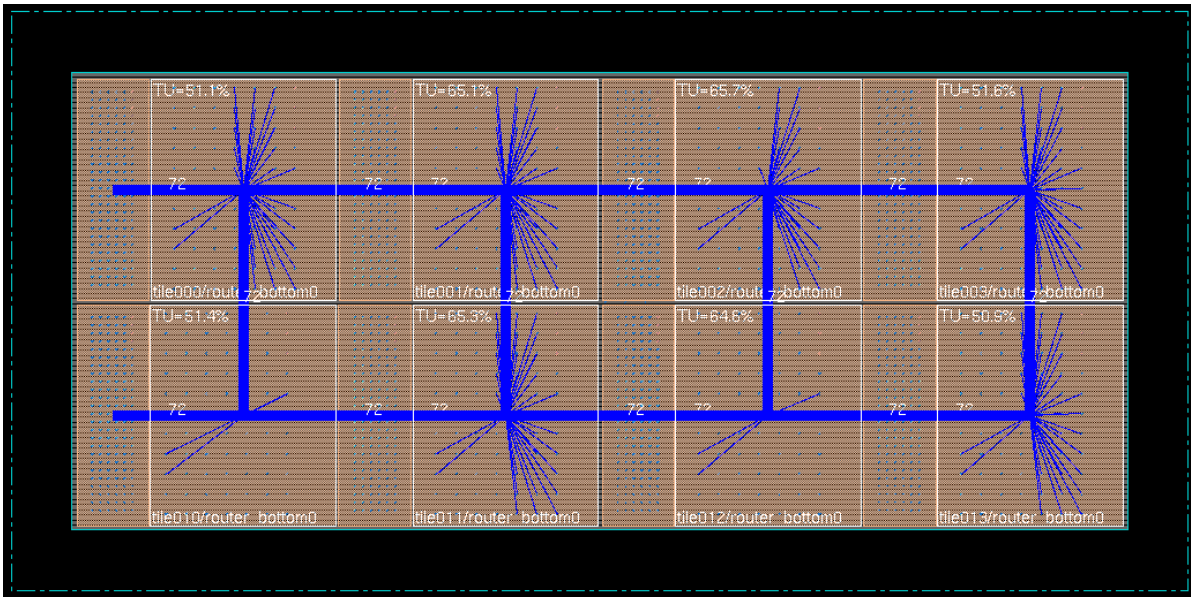


Figure 51: Floorplan of 3D Mesh NoC

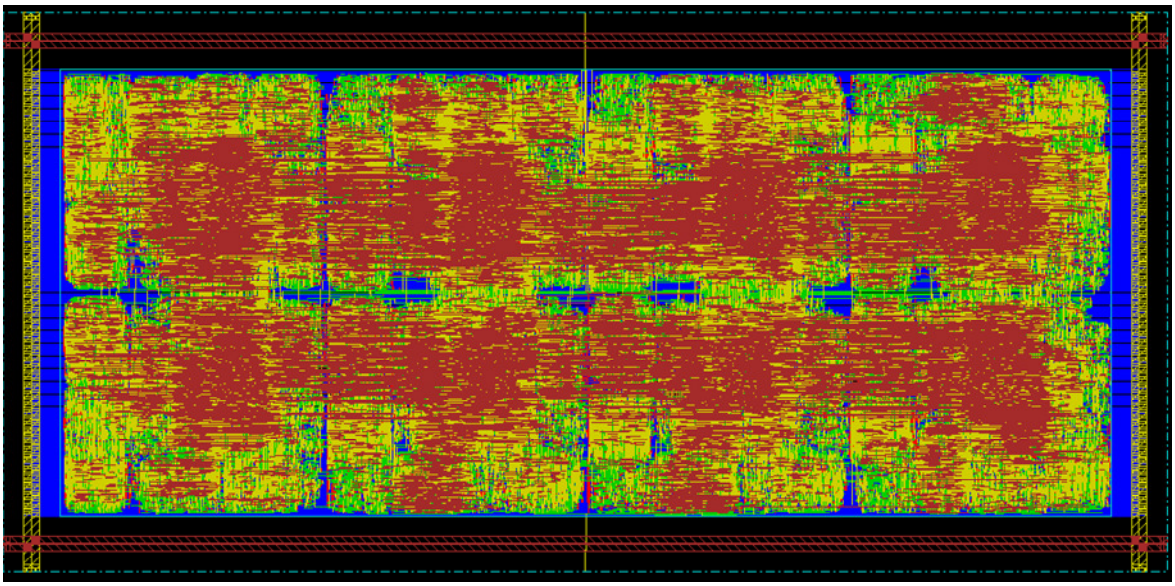


Figure 52: Routed layout of 3D Mesh NoC

5.4 3D NoC Architectures Exploration

In this section, we discuss several NoC architectures that are feasible to be designed using 3D stacking method. Several works have previously been performed for 3D partitioning analysis covering different target architectures such as muticore processor and memory architecture [158] [152]. Analysis of 3D NoC architecture based on physical design implementation is very limited in the literature where most of the reported works are using cycle accurate simulator [159] [151]. However, in this chapter, we present the analysis of 3D NoC architecture through physical design

implementation which provides more accurate performance analysis of 3D benefits and key parameters contributing to the 3D architecture performance is highlighted.

Previous work on performance evaluation of 3D NoC architecture through analytical formulation have shown that reducing tile area by partitioning it into several tiers will improve its power and latency [146]. Based on that analysis, we carried out further investigation by doing real physical design and implementation to obtain more realistic results and also to find what are the parameters affecting the performance of NoC architecture in 3D stacking. The stacking is done by simply dividing the datapath of the router to maintain homogeneous properties between the two tiers.

In this section, we explore two 3D NoC architectures through physical implementation to find the optimal solution benefiting from 3D technology. We only implement the NoC architecture (NIU and router) without processing elements because our focus is primarily on the 3D NoC architecture and also because we do not have a memory compiler 45 nm technology (to generate memory block for the processor) to be able to make a fair comparison with the 3D implementation of 130 nm technology. Even though one could argue that placement obstruction can be used to replace the processor components, we think that it would not be a good solution because these obstructions do not have routing characteristics as a normal logic block which will affect the overall routing characteristics and thus contribute to the other performance metrics.

5.4.1 3D NoC Partitioning

The FIFO buffers consume significant portion of silicon area in the NoC architecture (NIU and router). Thus, it is a good approach to partition it into two tiers and we have implemented this approach to partition the FIFO as well as other dapatpath components inside the NIU and router (such as crossbar) where the partition is done at bit-level. For example, for the 32 bit FIFO size, the resulting implementation will be 16 bits per tier. For the non-datapath components such routing logic, arbitration logic and FIFO control, we place it on each tier such the the area is balanced on both tiers. Figure 53 illustrates this partitioining method with respects to 2D and baseline 3D Mesh NoC architecture. Rather than using automatic tools such as HMetis to partition the design, we focus on dividing the datapath into two parts and place it into two tiers in order to preserve the homogeneous properties of tile block architecture between both tiers. Another reason for not using this automatic tool is because the tool also tries to optimize the nets between gates in the netlist with no capability of 3D placement meaning that logic cells can be interchangeably partition into the two tiers which will eventually affect the 3D timing path.

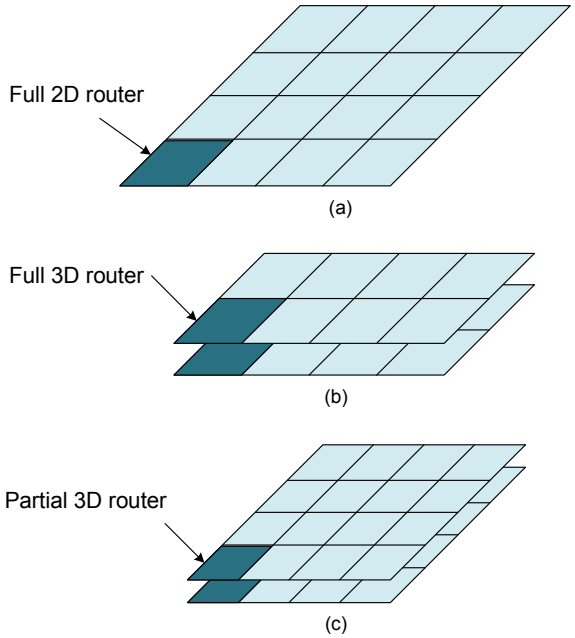


Figure 53: Partitioning method for the 3D NoC architecture (a) baseline 2D Mesh NoC (b) baseline 3D Mesh NoC (c) stacked 3D Mesh NoC

5.4.2 3DNoC1: 3D Stacked Mesh NoC

We have considered another approach for building 3D NoC architecture as depicted in Figure 54, Figure 55 and Figure 56. Previous work has shown analytically that for tile based multiprocessor architecture, the size of tiles play substantial part to the performance of 3D architecture [146]. Smaller processing element size will reduce the physical length of inter router links and thus improve NoC performance while reducing the NoC area only has little effects on the NoC performance. Therefore, rather than stacking the tiles on top of each other, instead we map the 3D NoC on the 2D layout and then partition it into two tiers. As shown in Figure 54, the green links represent logical vertical connections between 3D routers while the physical vertical links in orange color are basically the 2D logical links within the logic structure of NIU and router. By doing this, the area is slightly increased compared with the 3D Mesh NoC but reduced compared with its 2D architecture. However, this partitioning method requires higher number of inter-tier connections than 3D Mesh NoC. One disadvantage of this structure is that the inter-router wire links are not equal for all routers because vertical wire links are longer than other links.

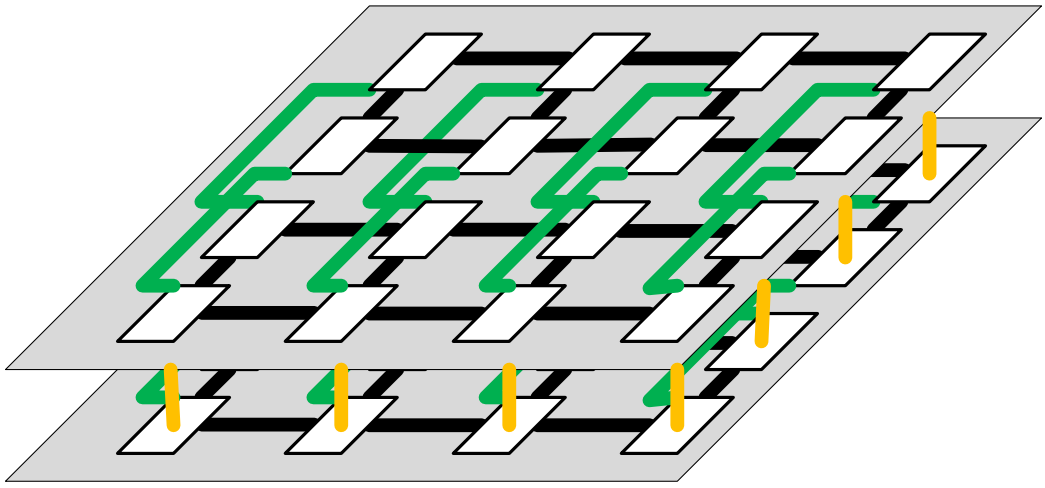


Figure 54: Block diagram of 3D Stacked Mesh NoC

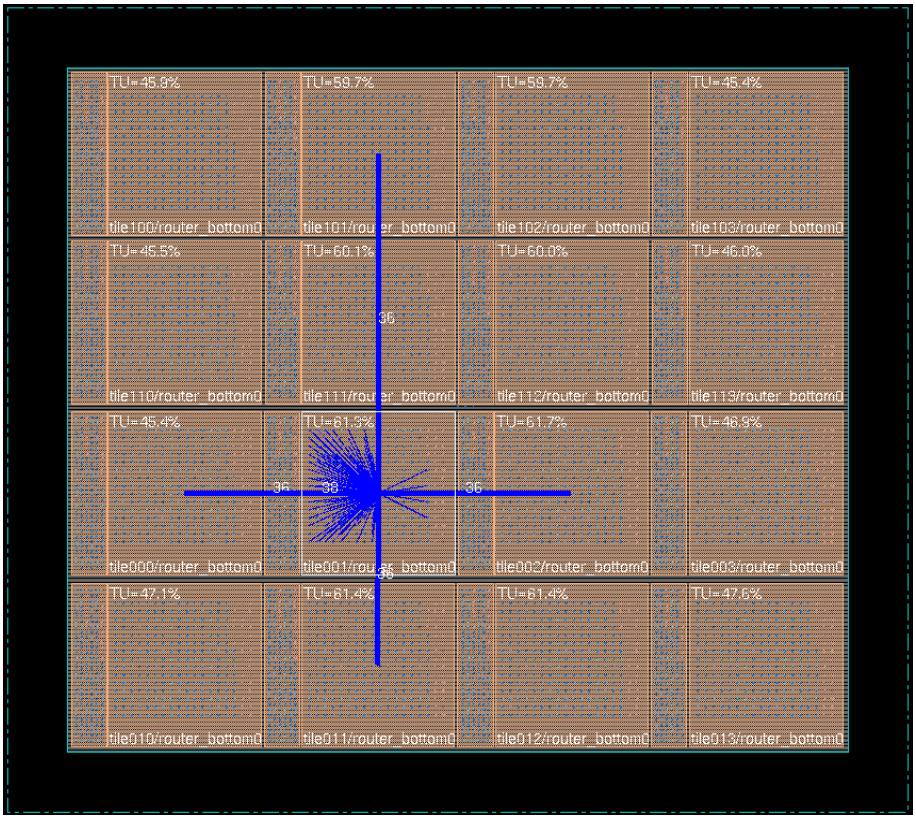


Figure 55: Floorplan of 3D Stacked Mesh NoC

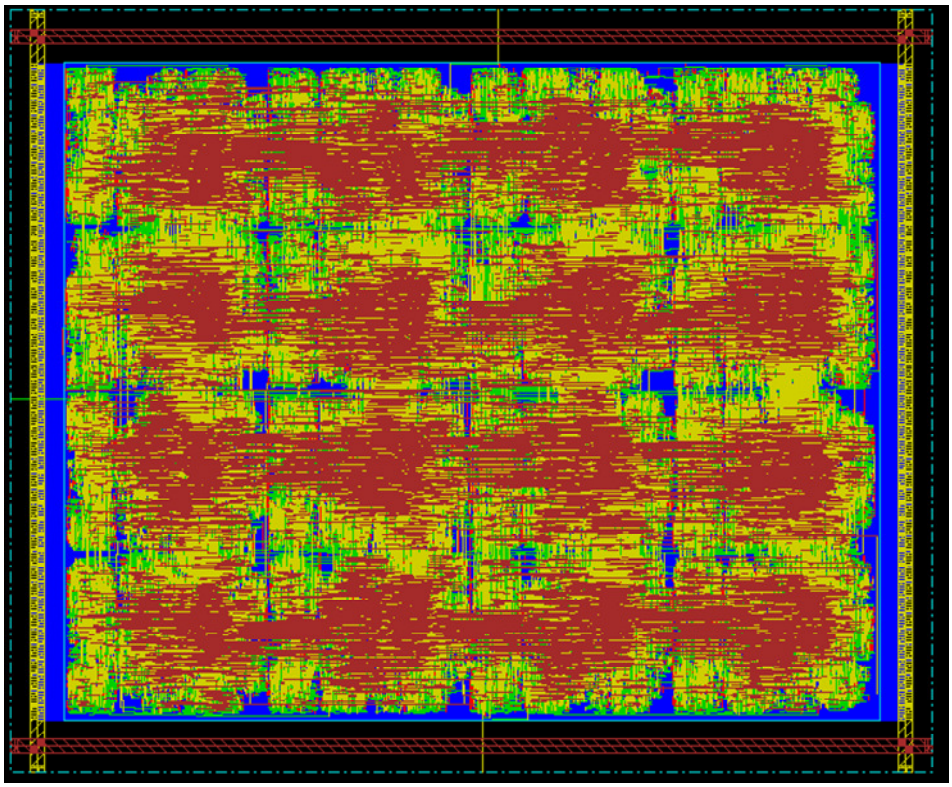


Figure 56: Routed layout of 3D Stacked Mesh NoC

5.4.3 3DNoC2: 3D Stacked Hexagonal NoC

Due to unequal inter-router wire links in the 3D Stacked Mesh NoC architecture because of the logical vertical links (green lines in Figure 54), to further optimize it, we proposed a new topology having same length of inter-router physical links called hexagonal topology shown in Figure 58, Figure 59 and Figure 60. Few works have studied this topology previously through analytical formulation indicating its advantages but lack of physical analysis from real designs and implementations [160]. As in the 3D stacked mesh NoC, the orange links represent 2D logical wires in the logic structure of NIU and router and is used to form the physical vertical connections between tiers. As we cannot floorplan the tile to create hexagonal area that has six edges with equal length using the current place and route tool, therefore we floorplan the tile by creating a rectangular area using the equation of $(a/2)^2 + b^2 = c^2$, where a is the tile's height, b is tile's width and c is the physical direct distance between the two tiles to determine the same optimal size of each tile. Compared with 3D Stacked Mesh NoC, this architecture consume slightly (shown in Figure 61 and Figure 63) more area due to the additional one more port in routers for diagonal connections. Although this topology has equal physical length of inter-router connections, there are empty areas due to the nature of this topology arrangement but it can be used for additional NoC structure such as monitoring infrastructure where the increasing monitoring capability will need

larger area to ensure reliable operation for the NoC. Recent work on dedicated NoC monitoring infrastructure showed that it consumed an additional area overhead of about 0.2% of total area which means that the available free area can be fully utilized [161]. In addition, the additional one more port in the router for the diagonal links as shown in the block diagram resulting increasing in area compared with 3D stacked mesh NoC architecture is due to the small structure used in this study.

5.4.3.1 Packet Routing

Routing logic must also be modified to support the diagonal directions. Consider the diagram illustrated in Figure 57. Basically the packets will be first routed through X direction and then to Y direction to reach the destination. However, in the case of a router with diagonal links, the packets will be routed through this diagonal links instead of Y axis link. Therefore, from the diagram, the packets will be routed from router 00 to 33 through router 11, 12 and 23. In order to differentiate which diagonal links should be used to route a packets, the routing logic will compare target router address with the current router address and then decide the direction of diagonal router link. For example, when a packet is in router 12, the router will compare it address with the destination router's address and therefore will select the diagonal links connected to router 23 instead of router 21. The diameter of the Hexagonal NoC can be formulated as $d = (x-1) + (y-1) - (x/2)$, where x is the number of hops in X axis and y is the number of hops in Y axis. As this type of routing is deterministic dimension ordered routing, thus it is a deadlock free routing. The general comparison between 2D mesh and 2D hexagonal topology is presented in Table 15.

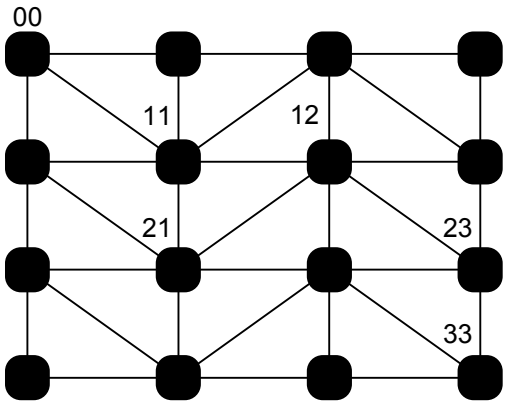


Figure 57: Routing method for hexagonal topology

Table 15: NoC topology comparison

Characteristics	3D Mesh NoC	2D Hexagonal NoC
Diameter	(X-1) + (Y-1) + (Z-1)	(X-1) + (Y-1) - (X/2)
Number of ports per router	6	6
Packet routing	XYZ direction	XY + diagonal direction

5.4.3.2 Physical Implementation

The physical size of the tiles is determined by measuring the distance between tiles such that the distance between each neighboring tiles is equal. This is to make sure that this rectangular floorplan is identical to the original hexagonal structure where it has six edges with the length of the edges equal with its radius. Although it is possible to create hexagonal floorplan using SoC Encounter, it is complex and thus we prefer a simpler solution to use rectangular floorplan. We adopted the equation of $(a/2)^2 + b^2 = c^2$, where a is the tile's height, b is tile's width and c is the physical direct distance between the two tiles to determine the size of each tile. We first fix the value of a and then find the value of b such that c is equal to a at the same time meeting initial target utilization. We also have derived mathematical formulation proving that the surface area of the square floorplan is identical to the original hexagonal structure. Let's say a equal to 579 µm, following the equation above will obtain the value of b equal to 500 µm and c equal to 577 µm with the initial target utilization of 60%. To compare the diameter for both topologies, consider an example of a 4 x 5 NoC. The diameter for 3D Mesh NoC is 6 while for 2D Hexagonal NoC is 5 and this reflect the benefits of hexagonal topology although different number of routers will result in different diameter for both NoC. Therefore, considering the topology aspect as well as their physical area, both topologies have almost equal design cost.

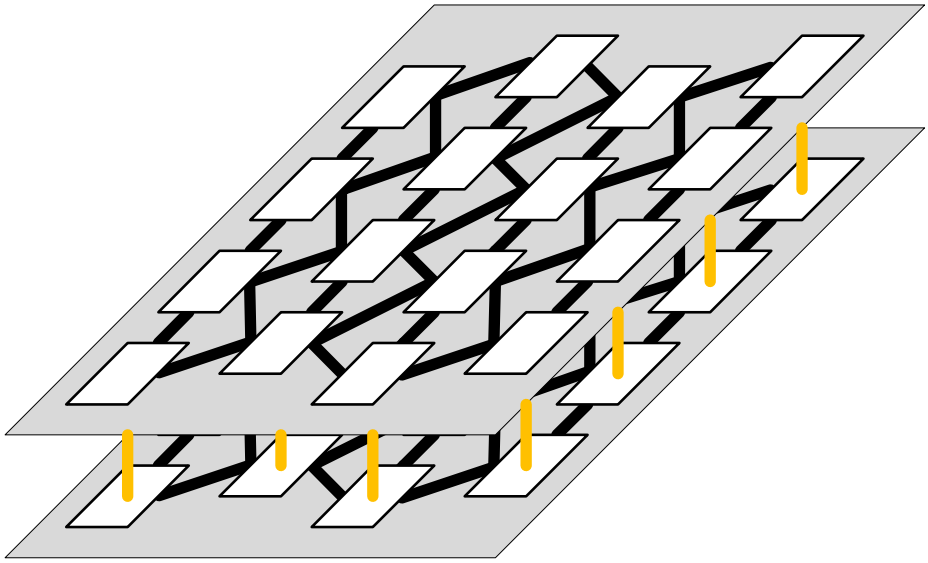


Figure 58: Block diagram of 3D Stacked Hexagonal NoC

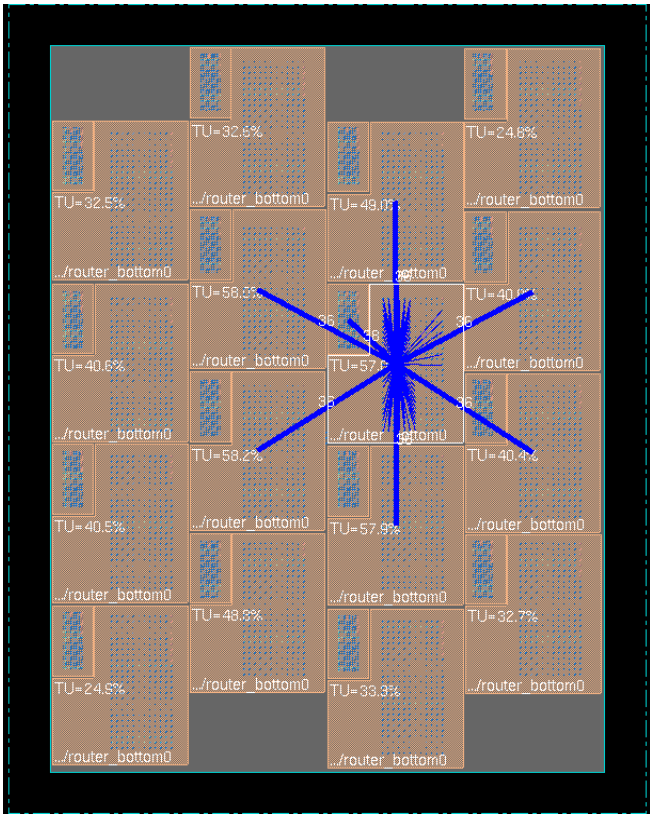


Figure 59: Floorplan of 3D Stacked Hexagonal NoC

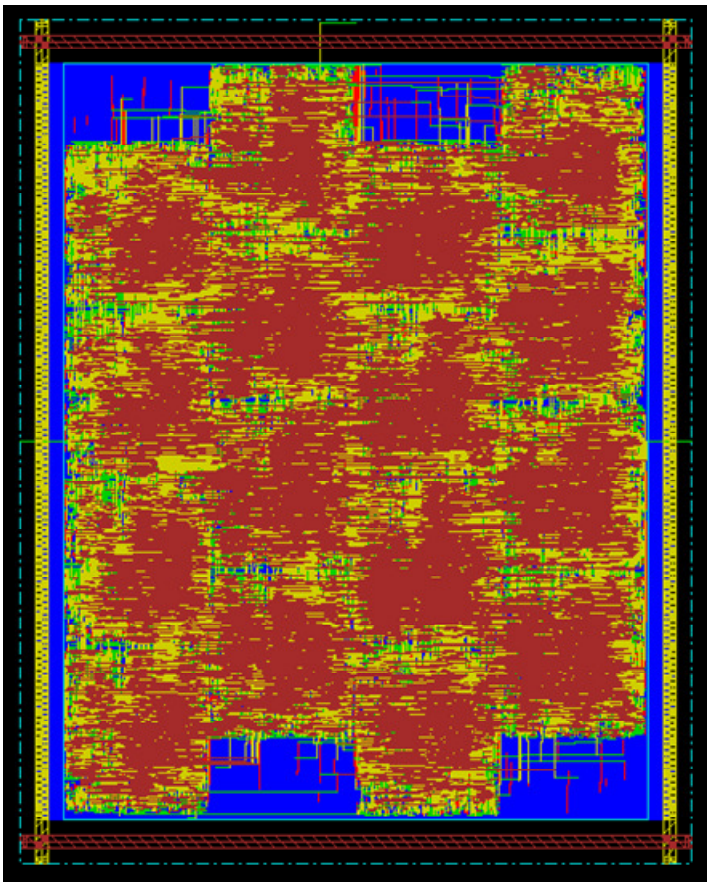


Figure 60: Routed layout of 3D Stacked Hexagonal NoC

5.5 Experimental Results

For older technology such as 130 nm and above, wire length effect is not significant and the delay in the critical paths is mostly determined by the delay of the gates and not wire delay. Due to that reason, 3D architecture provides little or no performance benefits over 2D design. As shown in Table 16 and Figure 61, the 3D NoC architectures do not benefits the speed and power consumption. The power consumption is higher in 3D architectures due to the additional gates as well as increased wirelength. Using the most advanced technology nodes together with the improved TSV dimension would provide higher performance benefit due to the fact that interconnect wirelength have strong impact on the global delay [162]. In this study, we used simple partitioning method to partition the 2D design into two-tier 3D architecture and results show that 3D does not improve performance of 2D design. However, some studies have shown that by using automatic partitioning tool could provide performance improvement over 2D architecture using old technology such as 130 nm and 180 nm [154] [163]. Partitioning is very important in 3D design primarily for old technology. Using automatic partitioning tool such as hMetis help to improve the performance of 3D architecture but it is still far from significant improvement. This is because the

tool tries to optimize the connections between gates in the synthesize netlist but does not able to perform in-place 3D optimization during place and route as in normal 2D optimization. At 45 nm, automatic partitioning tools provide higher performance improvement of 3D architecture than at old technology.

Table 16: Performance comparison of 3D NoC architectures in 130 nm technology

Parameters	3D Mesh NoC	3D Stacked Mesh NoC	3D Stacked Hexagonal NoC
Vertical connections per tier	1763	6261	7255
Number of links per router	370	370	444
Core area (mm^2)	3.24	4.37	5.43
Total wirelength (m)	12.48	14.01	17.03
Number of gates	295,956	295,510	338,337
Longest path delay (ns)	4.20	4.60	4.73
Power consumption @ 333 MHz (W)	1.44	1.25	1.40

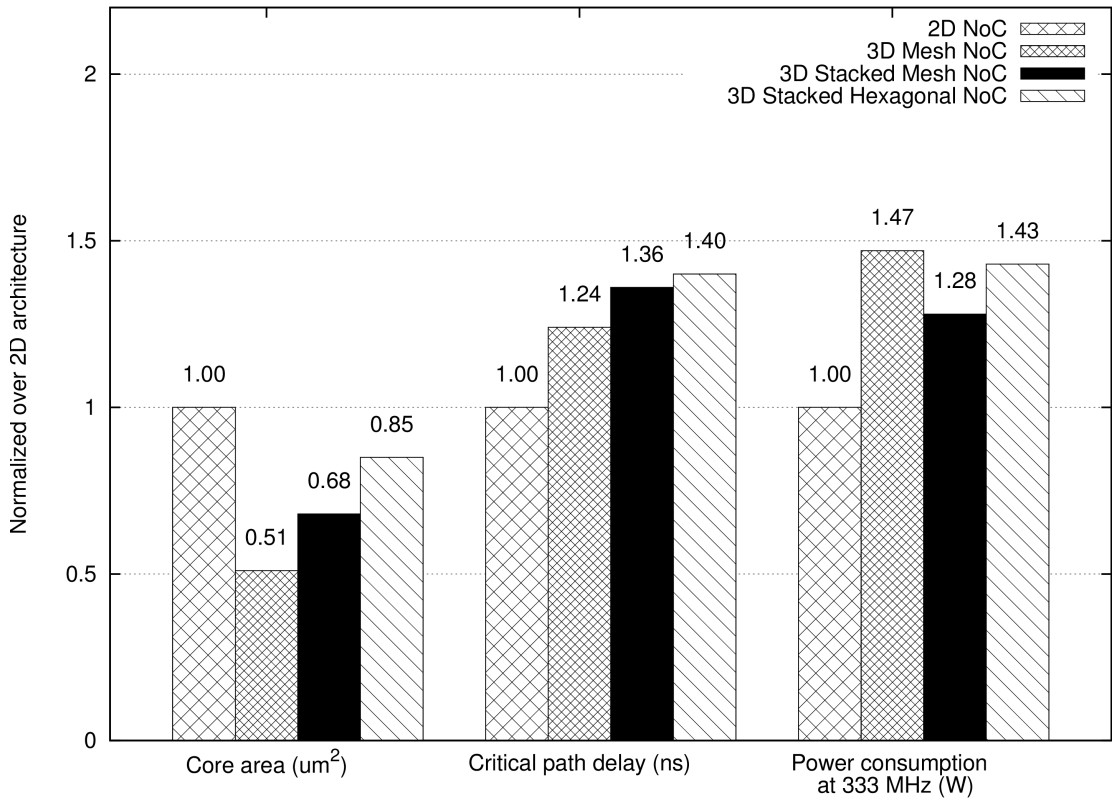


Figure 61: Performance comparison of 3D NoC architectures over 2D NoC in 130 nm technology

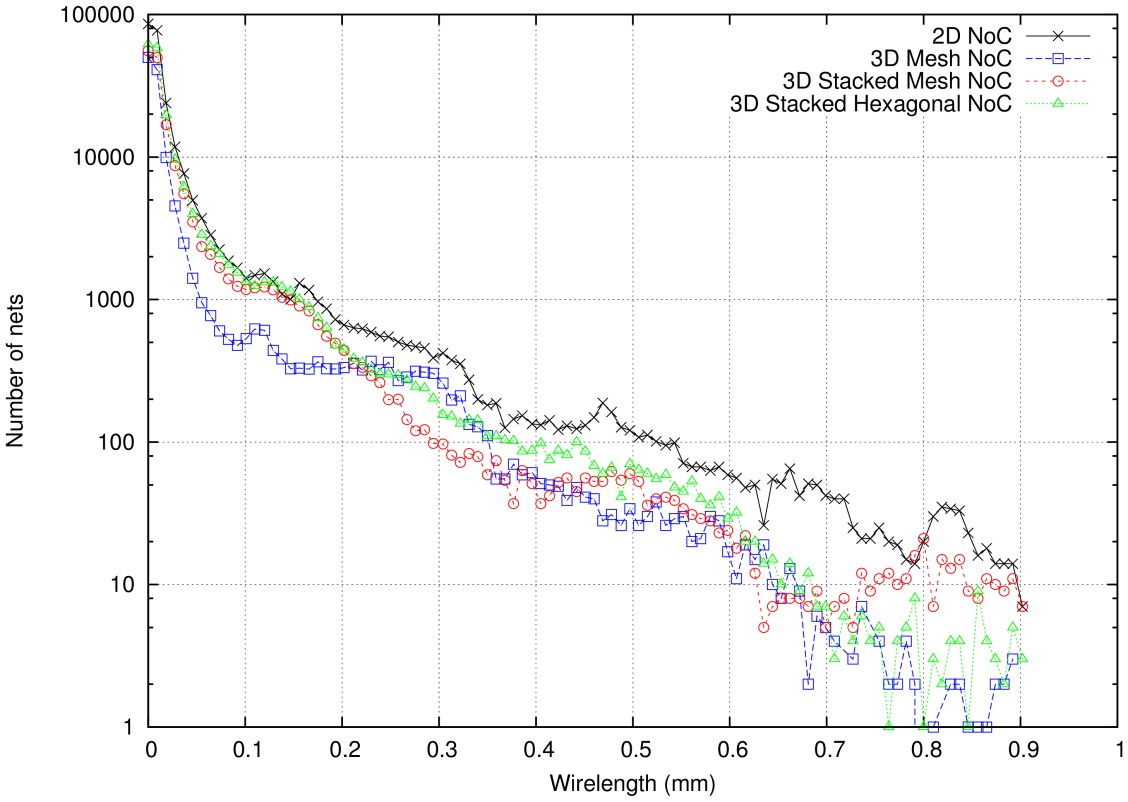


Figure 62: Horizontal wirelength distribution for NoC architectures in 130 nm technology

Table 17: Performance comparison of 3D NoC architectures in 45 nm technology

Parameters	3D Mesh NoC	3D Stacked Mesh NoC	3D Stacked Hexagonal NoC
Vertical connections per tier	1763	6261	7255
Number of links per router	370	370	444
Core area (mm^2)	0.79	0.91	1.01
Total wirelength (m)	5.5	5.9	6.5
Number of gates	255,088	257,220	290,896
Longest path delay (ns)	3.23	3.33	3.59
Power consumption @ 333 MHz (W)	0.23	0.24	0.26

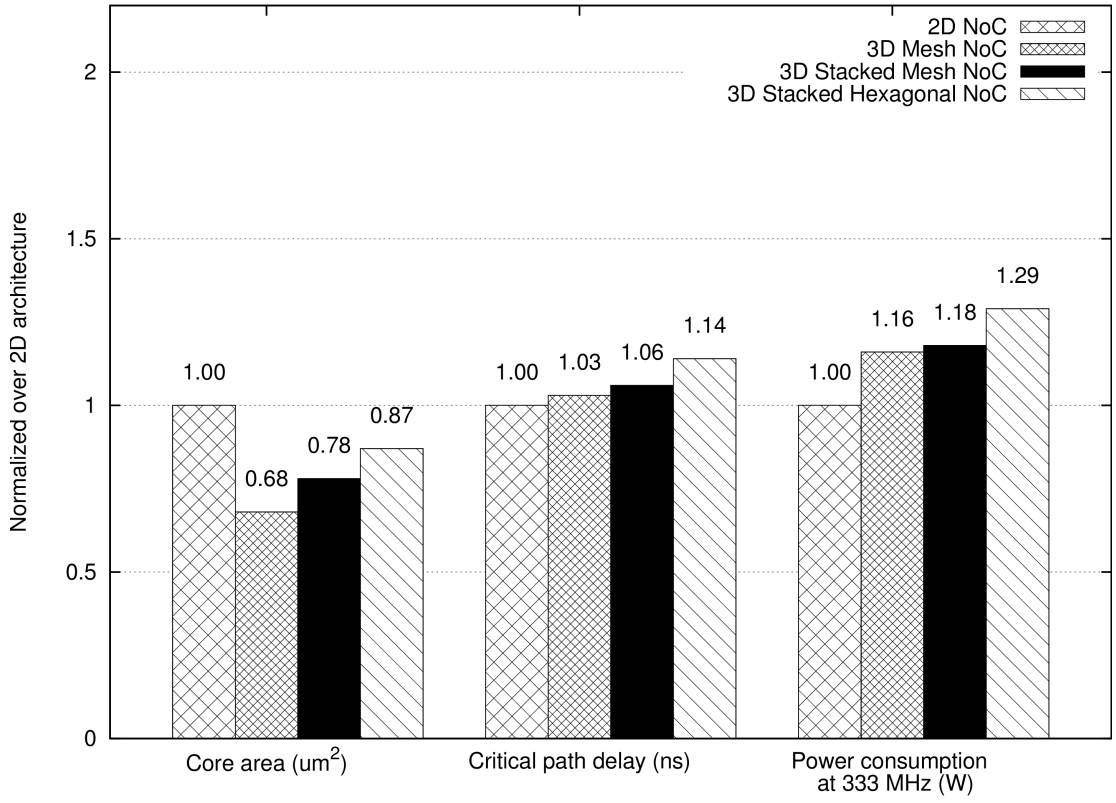


Figure 63: Performance comparison of 3D NoC architectures over 2D NoC in 45 nm technology

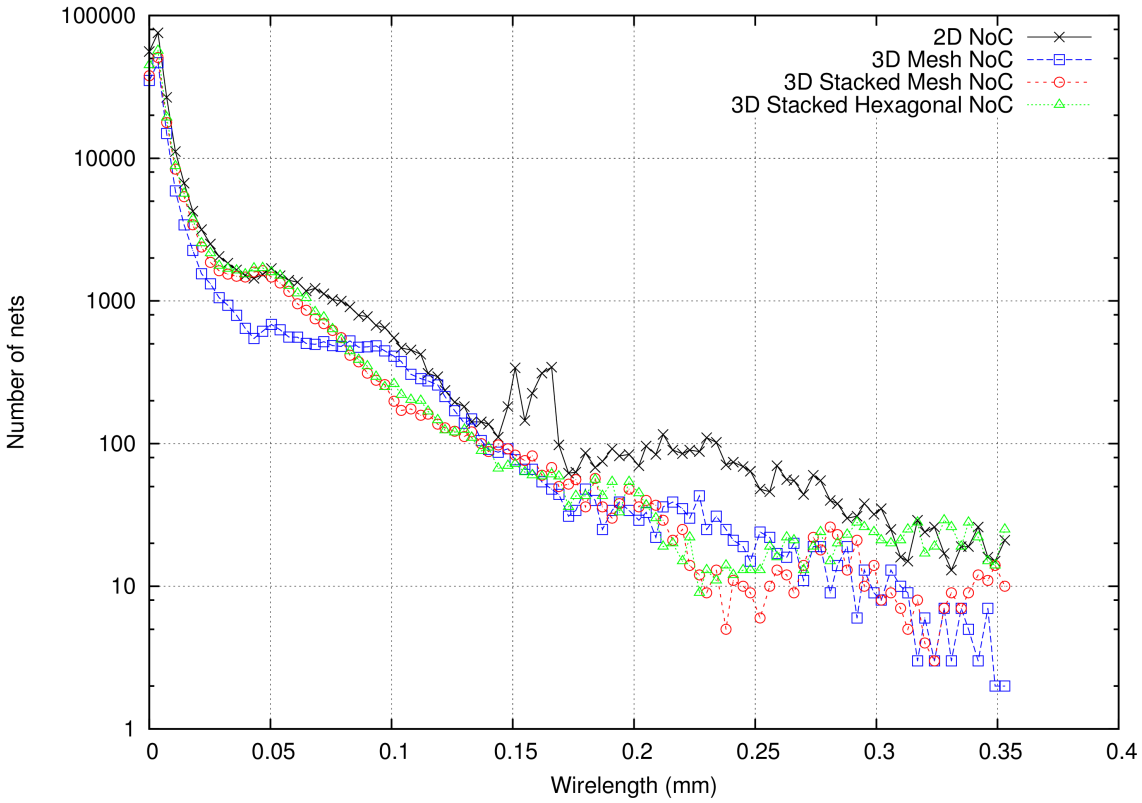


Figure 64: Horizontal wirelength distribution for NoC architecture in 45 nm technology

5.5.1 Wirelength Analysis

From the wirelength result in Table 16 and Table 17, we can see that total wire length for 3D architectures is increased compared with 2D architecture. The reason could be that we performed separate place and route step for each tier and the tool does not able to perform 3D architecture optimization. Compared with previous studies that used automatic partitioning tool, the total 3D wirelength is reduced [154] [43]. Automatic partitioning tool can be used to optimize the wire connections between gates prior to run place and route. Furthermore, the increasing number of gates in 3D architecture is also observed in the previous work due to the same reason as in the wirelength [152]. In our work, we use simple manual partitioning method and there is no wire optimization of the synthesized netlist prior to run place and route which increases the wirelength and the number of gates and thus contribute to the increased of power consumption in 3D architectures compared with 2D archictecture.

Horizontal wirelength distribution for the NoC architectures is shown in Figure 62 and Figure 64 for 130 nm and 45 nm technology respectively. This wirelength result is reported after the designs have been routed. It is shown that 3D NoC architectures generally reduce the horizontal wirelength distribution compared with 2D NoC architecture implementation. Wirelength reduction in 3D Mesh

NoC and 3D Stacked Hexagonal NoC is better than 3D Stacked Mesh NoC due to the equal inter-router wire links in this topology. For 45 nm technology, the graph shows that 3D Mesh NoC has better wirelength distribution than the other NoC architectures compared with 2D NoC architecture. 3D Stacked Hexagonal NoC has almost similar wirelength distribution to the 2D NoC architecture. As the area of the NoC architectures in this technology is`quite small where we can see that the longest wirelength is about 0.3 mm, therefore it is less accurate to evaluate the wirelength distribution of NoC architectures in this study. In addition, the 3D NoC architectures do not improve the longest wire compared with the 2D architecture in both 130 nm and 45 nm technology. Implementing larger architectures will have significant impact on the wirelength distribution of this 3D NoC architectures as have been demonstrated in [164].

5.5.2 Impact of Wire Delay

As for wire delay, older technology nodes (such as 130 nm) do not have significant wire delay effect to the speed performance. The critical path for all designs in this study is located within the tile block (from bottom tier to top tier in 3D stacked architecture) except for 3D Mesh NoC architecture where its critical path is between two routers. Looking at the 3D critical paths for all 3D NoC architectures in this study, the ratio of wire delay is about 3% of the total critical path delay. For comparison, the wire delay in 2D architecture using the same process technology is about 5.7% of the total critical path delay and thus we can generally conclude that 3D architecture in this technology will not offer any benefit in terms of speed. However, there is still an opportunity to gain benefit from 3D architecture by optimizing partitioning method as demonstrated by several works previously using older technology nodes although the results is not very significant compared with ideal improvement we should get [134] [65] [154]. Additionally, analyzing the critical path delay for the 2D architecture using 45 nm technology indicates that wire delay is about 1% due to very small area. We expect to see larger portion of wire delay in the critical path for 2D architecture with larger design.

The difficulties of optimizing 3D architecture using 2D EDA tools is that we are not able to directly optimize 3D critical paths through 3D timing constraints. Design flow based on 2D hierarchical design to optimize 3D architecture is not sufficient enough to gain maximum benefit from 3D architecture. 3D aware physical design tool, in particular 3D placement, 3D clock tree synthesis and 3D routing is very much needed to enable design space exploration to find 3D specific optimization for different type of architectures. Reducing wirelength due to area reduction in 3D architecture could help improving power consumption but does not necessarily improve 3D timing if the

wirelength is not in the critical path [165]. Although there are several works have been published related to timing aware 3D placement and 3D routing over the last few years [166] [167], it is not clear how much timing improvement is achieved compared with 2D architecture in order to assess the performance gain from 3D architecture. Optimizing clock tree buffer for the whole 3D architecture is also difficult using 2D place and route tool where clock tree contribute substantially to the delay of 3D critical path mainly for old process technologies.

5.5.3 Extrapolation of Physical Implementation Result

For designs using 45 nm used in this study, the 3D architectures still do not provide any improvement over its 2D design as shown in Table 17 and Figure 63. However, it shows a reduction trend of the gap between 3D and 2D architectures compared with the results in Figure 61 using 130 nm technology. If we look at the area, we can see that this design consume very small area (less than 1 μm^2) and this is the primary reason why there is no improvement obtained using 45 nm technology. Previous work have demonstrated for large designs (about 36 μm^2 in 2D architecture), substantial performance improvement (75% reduction in longest path delay) that could be achieved over 2D architecture using the same 45 nm technology because wirelength becomes significant [153]. Our design in 45 nm technology has relatively small area compared with the design in [153] in order to obtain significant performance improvement using only core level partitioning method (without partitioning tool). Table 18 shows the extrapolation of wire delay for 22 nm technology based on the wire delay obtain from a critical path in this study (results in 45 nm) and the data from ITRS 2007 interconnect report for intermediate wire. This extrapolation is intended to show that when the design used in this study is realistically large, we will see improvement for the proposed hexagonal NoC topology in stacked 3D architecture. The gate delay value for 22 nm is assumed to improve two times over 45 nm technology because it is two technology generations from 45 nm and the tile area (and thus the inter-router wire length) is assumes to be 3 mm x 3 mm for the 3D Mesh NoC considering the area of commercial grade LEON3 processor [153]. From the 3 mm inter-router wirelength of 3D Mesh NoC, we calculate the wire length for 2D Stacked Mesh and 3D Stacked Hexagonal using $x = 0.5\ a$ and $x = 0.53\ a$ respectively where a is the inter-router length for 3D Mesh NoC and x is the new inter-router length for each topology. The wire length for 3D Mesh and 3D Stacked Mesh is equal because 3D Stacked Mesh has half the area of 3D Mesh but has double the inter-router length for 3D logical vertical links. This extrapolation is simplified by ignoring the router area due to different number of ports for different topologies which is 4, 5 and 6 ports for 2D Mesh, 3D Mesh (also 3D Stacked Mesh) and 2D Hexagonal respectively. As can be seen from the table, the wire delay is becoming more significant for 16 nm technology and thus it will has strong

impact on the critical path delay especially for 3D Mesh NoC and 3D Stacked Mesh NoC (because of logical vertical links between routers) since it has longer inter-router wire links. The 2D Stacked Mesh NoC outperforms all the 3D NoC architecture in wire delay and eventually the total delay. However, the 3D Stacked Hexagonal NoC is better than 2D Stacked Mesh when considering the diameter because its routers have two more ports and the performance advantage will be higher when the network is large for example a 8 x 9 network.

Table 18: Extrapolation of delay for 3D NoC topologies using different process technologies and network diameter comparison

Technology / NoC architecture		Gate delay (ns)	Wire delay (ns)	Total delay (ns)	Diameter (8x9)
45 nm	2D Stacked Mesh NoC	2.6	1.5	4.1	15
	3D Mesh NoC	2.6	3.0	5.6	12
	3D Stacked Mesh NoC	2.6	3.0	5.6	12
	3D Stacked Hexagonal NoC	2.6	1.59	4.19	11
22 nm	2D Stacked Mesh NoC	1.3	4.95	7.55	15
	3D Mesh NoC	1.3	9.9	11.2	12
	3D Stacked Mesh NoC	1.3	9.9	11.2	12
	3D Stacked Hexagonal NoC	1.3	5.25	6.55	11
16 nm	2D Stacked Mesh NoC	0.6	8.85	9.45	15
	3D Mesh NoC	0.6	17.7	18.3	12
	3D Stacked Mesh NoC	0.6	17.7	18.3	12
	3D Stacked Hexagonal NoC	0.6	9.38	9.98	11

5.5.4 Impact of 3D IC design using 2D EDA Tools

Although we are able to design and implementing the 3D architecture using state-of-the-art commercial 2D EDA tools, the performance benefits we can get from 3D architecture is sub-optimal due to the limitation of the 2D tools that do not able to perform 3D-aware physical design especially doing 3D optimization between different tiers as the tools do not see the complete 3D architecture [168]. Using various additional tools such as 3D floorplan and 3D placement is also showed relatively small performance improvement in terms of power reduction and speed improvement hence does not justify additional cost for stacking to be able to adopt this technology [154]. On the other hand, using 2D EDA tools with complicated design flow together with fine-

grain partitioning method can gives pronounced performance improvement [128] but requires much design efforts and design time and also lack of design exploration capability to be able to conduct design trade-off before making any decisions. The true 3D-aware design tools with the capability of solving grand challenges in 3D physical designs [169] is mandatory to enable widely adoption of 3D IC technology in industry.

Specific to the Tezzaron 3D IC technology where dies are connected using microbumps, its small structure and pitch allow high vertical interconnection density with small/negligible delay but additional design efforts are still required in order to achieve significant performance improvement. Using manual assignment for the vertical signals does not able to provide compelling performance improvement due to the microbumps location that is not optimized relative to the cells in both tiers. Additionally, using multiple microbumps per signal could also gives noticeable improvement especially when integrating with the 3D partitioning methods as demonstrated in [153] which is not implemented in this work and subject for future research direction.

5.6 3D IC Implementation for MPSoC Architectures: Mesh and Butterfly NoC

In this section, we describe the architecture and implementation of two 3D MPSoC architectures based on different NoC topologies, one being developed by a team in ENSTA ParisTech in Paris and the other one by us in GIPSA-Lab in Grenoble. The designs have 16 processors communicating using a NoC and spread on two tiers are discussed in detail and are targeted to be fabricated using Tezzaron technology with 130 nm Global Foundries standard library. The purpose of this work is to accurately measure NoC performances in real 3D chip when running several applications to evaluate the impact of 3D MPSoC architectures when compared with its 2D implementation.

As discussed earlier in this thesis, simulation is one of the most common methods that have been used for performance evaluation of 3D architecture and there is a limited number of works doing performance analysis based on real implementation. Summary of 3D architecture implementations as explained in chapter two reveals many 3D chips have been taped out targeting various implementation objectives such as 3D SoC implementation, memory bandwidth improvement, fault tolerant techniques and modular multicore architecture. Yet, none of the previous fabricated chip considers a complete MPSoC with NoC architecture to study its performance when running applications which are the main goal of the work in this chapter.

The purpose of this work is to evaluate the 3D NoC performance when running applications in real

3D fabricated chips based on two-tier Tezzaron 3D technology as has been detailed in chapter two in this thesis. The designs will be sent for fabcrication through Chip Multi-Projet (CMP). Two different MPSoC architectures is being implemented based on Mesh NoC (will be called MPSoC1) implemetend by our group and Butterfly NoC topology (will be called MPSoC2) implemented by the ENSTA ParisTech team with the common Openfire processor in both architectures. Physical design implementation comparison of both MPSoC architectures will be discussed in the next sections.

5.6.1 3DMPSoC1: Mesh Topology

The first MPSoC architecture as shown in Figure 66 and Figure 68 has eight processors connected using 4x2 mesh NoC in each tier where the NoC is based on 3D routers. The floorplan of tile block in illustrated in Figure 65 and Figure 67 for top and bottom tier respectively. By not stacking the same memory block on top of each other, it could help to reduce the temperature of the 3D chip as studied in several works [170]. Connection between tiers is achieved by means of vertical ports of each router physically through microbumps structure. Summary of the physical design implementation is shown in Table 19. Total inter-tier connection is 594 connections (35 bits flit data + 2 tx/rx signals + 2 JTAG signals for one direction vertical port router). Synchronization between processors is implemented using FSL linked to the NoC separated from FSL used for data communication. A processor will synchronize before accessing its data memory by waiting for a tag word in its FSL sent by the writer processor. This is a simple synchronization hardware implementation in order to reduce die area.

We use an IEEE 1149.1 JTAG port for off-chip interface. The JTAG controller is located at the bottom tier and connected to the outside chip using TSV within the I/O pad. Loading instruction and data memory for each processor is also using the JTAG port. Also the data memory of one processor (id 0) is connected outside in order to have fast access to results and be able to provide new input data.

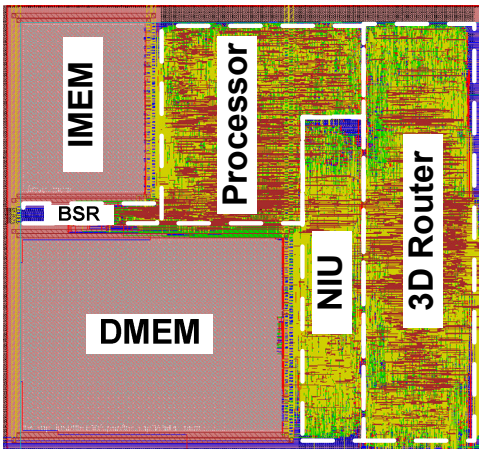


Figure 65: Tile block floorplan of 3D MPSoC1 (top tier)

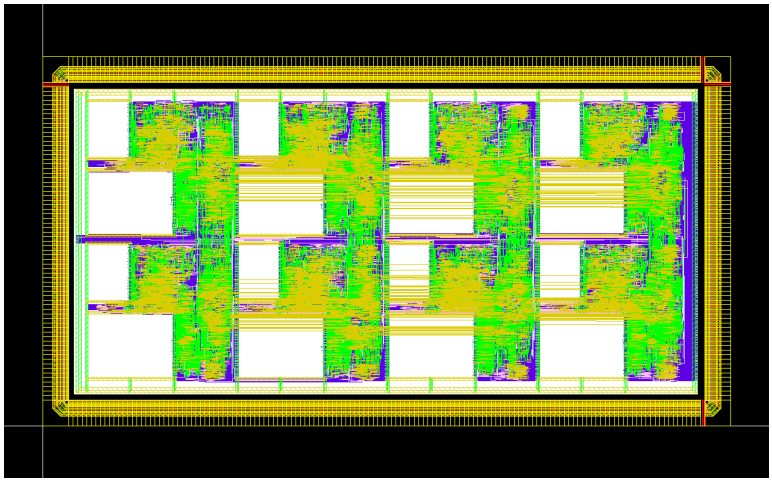


Figure 66: Virtuoso layout of 3D MPSoC1 (top tier)

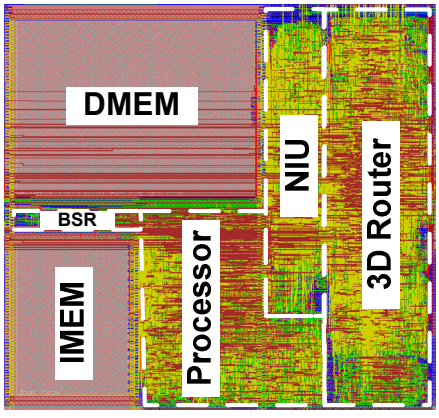


Figure 67: Tile block floorplan of 3D MPSoC1 (bottom tier)

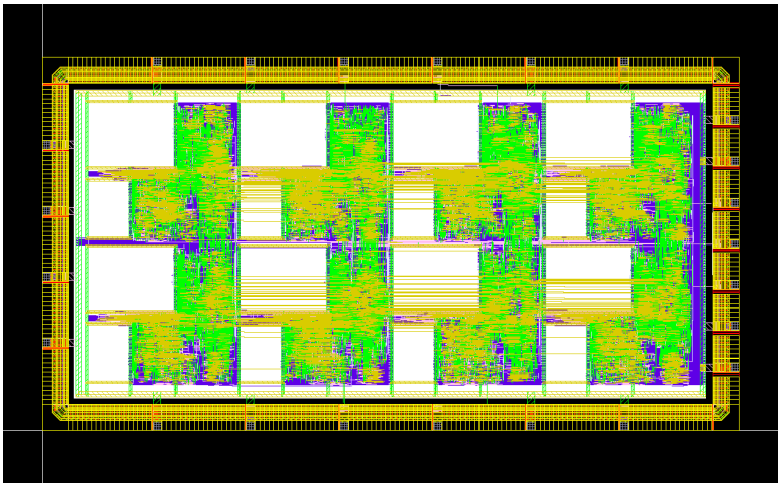


Figure 68: Virtuoso layout of 3D MPSoC1 (bottom tier)

Table 19: MPSoC1 physical design characterics

Parameters	Details
Number of gates per tier	1.3 million
Number of inter-tier for signals	594
Number of inter-tier for power/ground	5568
Number of IO signals	7
Off-chip interface	JTAG IEEE 1149.1
Target frequency	100 MHz for processor, 300 MHz for NoC
JTAG frequency	10 MHz

5.6.2 3DMPSoC2: Butterfly Topology

The second MPSoC architecture is based on a 2D 8x8 butterfly NoC topology linking the 8 master processors to the 8 slave memories as depicted in Figure 69 that has been developed by another team in ENSTA ParisTech in Paris. Two interface blocks which are FSL-to-OCP and OCP-to-NTTP have been used to allow communication between the processor and the NoC where the former has been design using VHDL while the later is generated automatically together with the NoC RTL files. The butterfly NoC architecture has three stages with four routers in each stage and the links from a stage to another have different lengths. With this topology we can have long links which can be ideally reduced with vertical connections in 3D Design. As in the MPSoC1

architecture, FSL Bus is used to connect the Openfire processor to the NoC via network interface with several custom interface adapters to be able to suit with the protocol supported by the NoC (NTTP). Due to the many FSL ports supported by this processor, we connect each processor to the NoC in the same layer with an FSL port 1 and keep the FSL port 2 to make a vertical link with the processor in the other tier as shown in Figure 70. With these vertical connections, processors from the top tier and bottom tier can communicate and synchronize together. Figure 71 shows the routed layout of bottom tier for this MPSoC architecture.

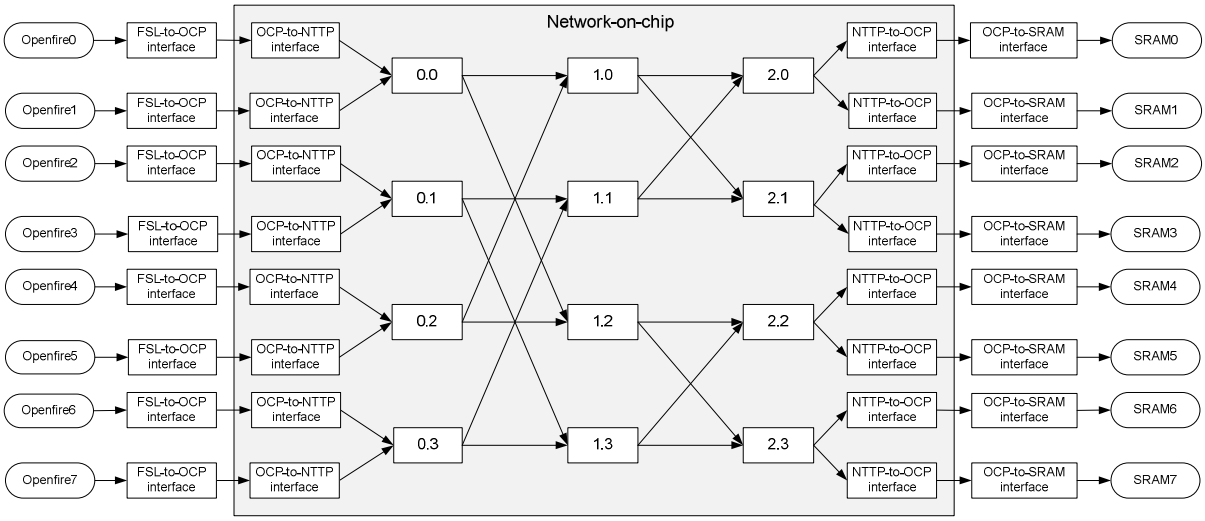


Figure 69: NoC block diagram for 2D MPSoC2 architecture

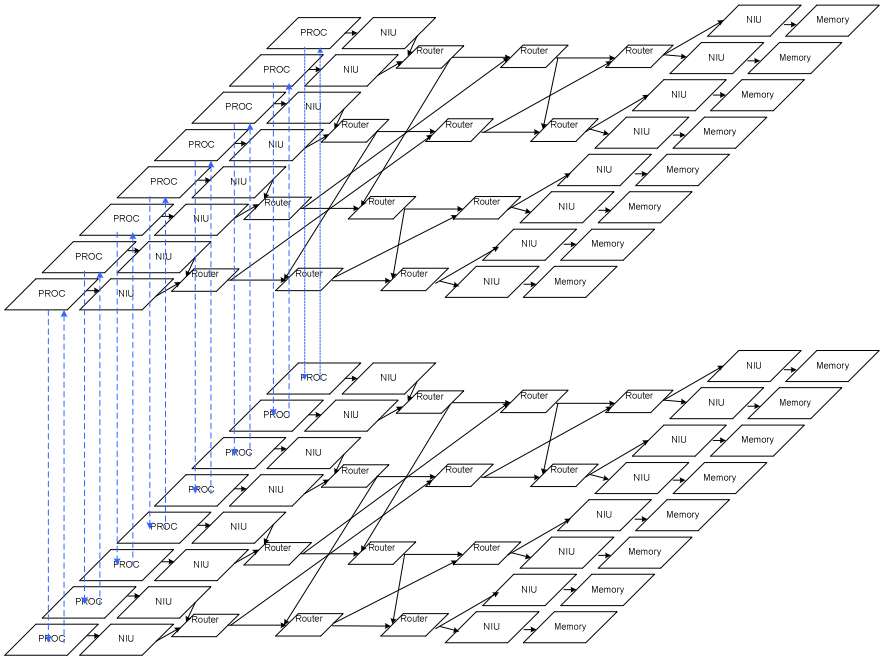


Figure 70: 3D MPSoC2 block diagram

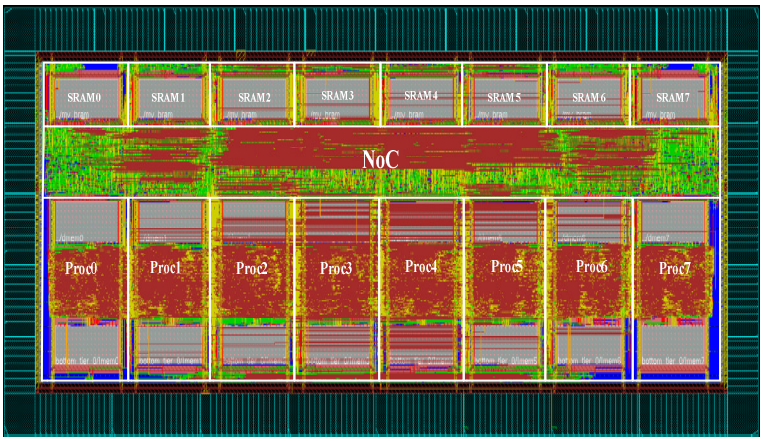


Figure 71: Routed layout of MPSoC2 architecture (bottom tier)

5.6.3 3D MPSoC Implementations Comparison

The physical design implementation comparison of both MPSoC architectures is shown in Table 20 where we can see that the number of vertical interconnection is not that much different between the two architectures. In addition, MPSoC1 architecture which is based on mesh NoC topology has symmetry architecture from the NoC topology point of view while it is not the case for the MPSoC2 using butterfly NoC. Analyzing the performance of both architectures on real 3D implementation when running several applications will be provide more accurate results on how the NoC topologies benefit from the 3D architecture.

Table 20: 3D MPSoC implementations comparison

Parameters	MPSoC1	MPSoC2
Number of PE	16	16
Number of shared memories	-	16
NoC topology	Mesh	Butterfly
Number of tiers	2	2
Number of vertical connections	594	560
Symmetry (NoC topology)	Yes	No
Die size (mm)	3.2 x 4.89	1.99 x 4.95
Operating frequency (MHz)	100 (processor), 333 (NoC)	100

Conclusion

We have described performance analysis of 3D NoC architectures through physical design implementations in order to accurately analyze and to have better understanding of benefits of NoC architectures in 3D technology. We used different standard libraries to evaluate its impact on the performance of 3D architecture and investigate the role of several parameters with regards to physical design property such as wire delay to relate to the 3D architecture performance. The stacked 2D NoC topology (refrerring to the 3D Stacked Hexagonal NoC) is found to have better performance compared with the 3D NoC topology (3D Mesh NoC) in advanced technology. Summarizing the results, partitioning is shown to be a very important step for designing 3D NoC architecture and 3D-aware tool is also a must for physical design implementation with the ability to perform 3D optimization for various design objectives as in today's 2D tools. It is also very much needed in order to fully obtain the advantage of 3D technology. In addition, using automatic partitioning tools as well as 3D-aware physical design tools will improve the 3D architecture performance using both old process technologies as well as advanced process technologies but it will be more significant in advanced technologies as wire delay is dominant when designs are sufficiently large. We have also compared in this chapter the physical design results of two different MPSoCs implemented using 3D Tezzaron Technology. The implementation of MPSoC1 is based on mesh NoC topology while the MPSoC2 uses butterfly NoC topology. For MPSoC1, with its short links of inter-router connection and its symmetric architecture, this topology could not really demonstrate the benefit of 3D implementation compared with the NoC topology in MPSoC2 architecture that has unequal inter-route wire links.

CHAPTER 6
HETEROGENEOUS STACKING OF 3D NOC-BASED MPSOC ARCHITECTURE

6.1 Introduction

We have described the implementation of homogeneous stacking in 3D architecture in the previous chapter where we divived a 2D architecture into two identical layers and stacked them using microbumps. In this chapter, we explore another possible stacking approach for the 2D architecture using heterogeneous 3D stacking and compare its performance with its 2D architecture. In this experiment, we refer to heterogenous 3D stacking as stacking of different architectures in different layers for example memory layer on top of logic layer using same process technology. To enable this, we implement a complete multiprocessor with NoC architecture in this study to analyze 3D architectural design issues through physical design implementations.

Although 3D IC technology can be used for various complex SoCs (System-on-Chip), prime candidates are MPSoC. Several MPSoC architectures have been designed [171] for complex applications and the trend towards manycore [172] is pushing for 3D implementations. First 3D implementation of MPSoC have been recently reported [83]. Due to their modular architecture, MPSoC offers several opportunities for physical implementation optimizations among them clocking and power consumption through Globally Asynchronous Locally Synchronous (GALS) [2]. Under tight application specific constraints such as power and silicon area budget, heterogeneous MPSoC will be the key consideration to achieve desired performance [173]. MPSoC have systematically adopted Network on Chip (NoC) for inter-IP communications and the benefits of NoC in all devices technologies (ASIC, FPGA) are now well established and demonstrated [142] [174]. Employing NoC architecture allows designers to apply various techniques such as voltage frequency island (VFI) and power aware applications mapping to achieve different design objectives [175].

3D IC technology opens up new opportunities for architecting multiprocessor architecture to achieve desirable target performance. As there is no real 3D-aware tool available to date, therefore the common approach to perform architectural exploration is by determining how to partition the design into several stacks using software/scripting algorithm separately from the EDA tools or otherwise doing it manually. The motivation of study in this chapter is to evaluate the performance of heterogeneous 3D architecture for GALS-based multiprocessor architecture because dividing the communication architecture from the computation architecture physically in different physical

layers gives opportunities for power and thermal optimization. Besides, the lack of physical design studies of heterogeneous 3D stacking especially for GALS style implementations are also another point of interest to conduct this study. GALS architecture allows different blocks operating at separate voltage and frequency making it ideal for dynamic power management system and also enables easier IP integration having different clock domains [176].

In this work, we explore the heterogeneous 3D MPSoC stacking architecture and evaluate its timing and power consumption compared with its 2D counterpart. Moreover, we also conduct an experiment to show that microbumps pitch is an important parameter need to be taken into consideration when doing planning of the 3D architecture such as partitioning and floorplanning although it does not have drawbacks of routing blockage and large keep-out-zone as in TSVs. Physical design implementations have been performed by varying the pitch of the microbumps for the vertical interconnection of logic and memory block to investigate the effect on the 3D architecture performance.

6.2 Related Works

3D heterogeneous architectures have been studied by several researchers but mostly restricted to analysis from software simulation. The most common approach to implement heterogeneous 3D stacking is using memory on logic stacking primarily to achieve higher memory bandiwth due to advantage of huge amount of vertical inconnections. In [124], the have designed and implemented memory on logic architecture for the 64 multicore processor where each data memory for each core is place on another layer on top of its logic layer. The instruction memory is placed on the logic layer in order to have maximum size for data memory for each core. To achieve maximum memory bandwith, the processor core is designed specifically to consume memory bandwith at every cycle from the 3D stacked memory by allocating one slot for the memory instruction. However, they do not use NoC architecture for the communication architecture due to the stable, predictable and regular communication pattern in their data-parallel application. Instead, they use buffer-based architecture to allow processors communicate between its neighbouring blocks. In [177], heterogeneous memory-on-memory architecture is studied by stacking SRAM cache with logic on the 3D DRAM layer with the aim to optimize both performance and energy efficiency. By folding the DRAM bank layers into 4 layers and then share the same TSVs bus to the logic layers, it reduces the energy from transferring entire row signals. Another work on heterogeneous stacking is done by [178] where they stacked heterogeneous DRAM layers on processor layers. Performance analysis is done using software simulation based on modified CACTI and M5 simulators for full

system simulation with multicore processor.

With regards to 3D architecture using NoC, we found limited number of works about heterogeneous stacking based on NoC architecture especially the one implementing physical design. In [144], 3D architecture using combination of heterogeneous IP cores layer and homogeneous mesh NoC layer is studied and performance analysis is done using cycle accurate simulation. The main reason behind their work is that heterogeneous multicore architecture does not have the same IP core and thus the different size between each IP core makes it not suitable to use Mesh NoC where it is normally based on homogeneous multicore architecture with same IP core size. In order to use mesh NoC with the heterogeneous IP core architecture because of regular properties of mesh topology, 3D architecture can be used to realize it by stacking both different layers on top of each other. Another work in [179], they presented a three tiers heterogeneous architecture by using a VesFET-transistor based NoC architecture in the middle layer between core and cache layers in order to reduce the router to router wire links compared with 2D and normal 3D implementation. Their analysis based on HSPICE simulation shows power and latency improvement basically because of router to router distance reduction.

State of the art electronic design usually facilitates GALS architecture to be able to meet design specifications especially for tight power requirements. Power consumption can be reduced up to two times lower for the same architecture using fully synchronous implementation at smaller area overhead using fine-grained clock domain partitioning [180]. Multiprocessor implementation with NoC architecture is nicely fitted with the GALS style where communication architecture can be separated from the computation architecture with different clock speeds hence enabling high performance system with power efficiency [181]. To the best of our knowledge, there is no work investigating the implementation of GALS style 3D multiprocessor architecture to date wherein the main motivation of this study. Deploying GALS architecture in 3D IC tehnology is also very exciting due to fact that it gives more design space to be explored with the existence of the vertical architecture in meeting various target implementation requirements.

In this work, we based upon the work in [144] to further investigate the performance of heteregoneous stacking for NoC-based multiprocessor architecture with slight modification to be more realistic implementation considering the router and processor area from the fabricated designs. In particular, a part of the processor component is placed in the same layer with the NoC architecture to cover the empty area due to the smaller NoC area than the processor which will be more detailed later in this chapter. Using Tezzaron two-tier technology, we carried out physical

design implementation of the heterogeneous 3D stacking MPSoC architecture and compare its performance with the 2D architecture from architectural point of view. This study provides additional architectural exploration for the previously done homogeneous stacking of 3D NoC architectures as well as architectural exploration of the GALS style implementation in 3D architecture. Deep understanding about how performance is affected by different 3D architecture implementations is essential to find the right architectural candidate to fully benefit from the 3D technology.

6.3 Baseline 2D NoC-based MPSoC Architecture

6.3.1 Processor Architecture

We use an open source processor for our implementation which is readily available without spending much time to develop a new processor. The Openfire processor as shown in Figure 72 and Figure 73, is downloaded from Opencores.org. It is a Microblaze clone which is based 32-bit Reduced Instruction Set Computing (RISC) architecture using Harvard architecture that supports Microblaze instruction set architecture (ISA) and compiler tool chain [182]. Comparing with MicroBlaze processor that has hardware multiplier, hardware divider, barrel shifter and floating point unit, Openfire processor has only hardware multiplier and also supports On-chip Processor Bus (OPB) for external interface particularly for accessing instruction and data memory. Although there are other open source synthesizable Microblaze clones available to be used [183], we choose Openfire because it has Fast Simplex Links (FSL) ports (basically a FIFO that support dual clock domains) that we need for simple data and synchronization communication between processors and NoC rather than using more complex interface such as Open Core Protocol (OCP) and Advanced eXtensible Interface (AXI) which require complex logic for implementation. It supports up to 16 FSL ports as in MicroBlaze allowing us to integrate additional functions such as NoC monitoring service using simple interface to the processor.

The Openfire processor is a simple processor developed initially for configurable processor research [184] but have been used for other purpose [185]. Thus, because of its simplicity, it will not require a large silicon area and thus can be used to develop any small application for testing the NoC in 3D architecture. Additionally, we use only 4 KB for instruction and 4KB for data memory in order to limit the die area. These memories are generated using Artisan memory compiler. The processor has 32-words register file implementing using flip-flop registers which consuming most of the processor's logic area.

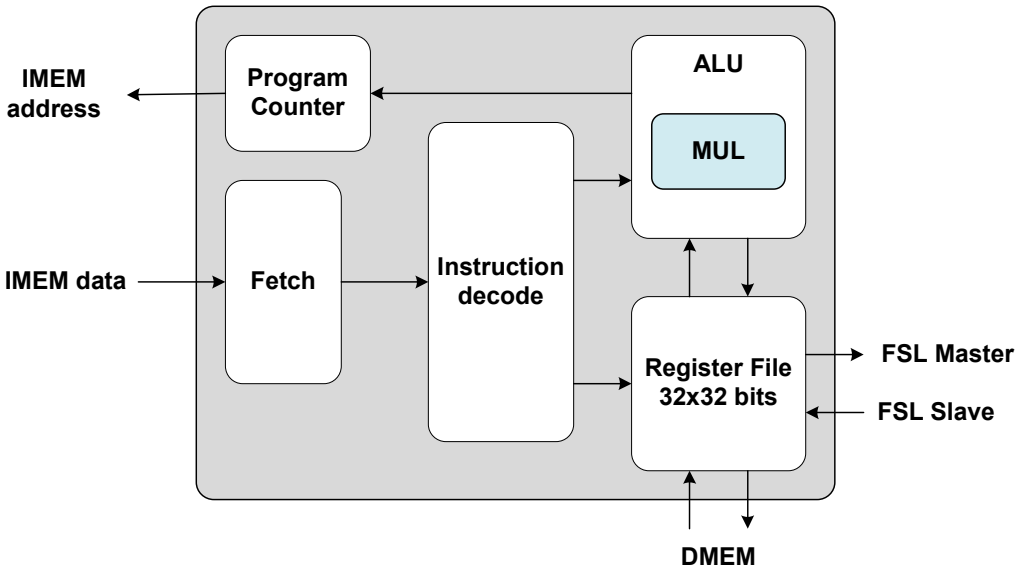


Figure 72: Openfire processor block diagram

6.3.2 NoC Architecture

The NoC architecture in this experiment is based on 2D mesh topology implemented using router and network interface architecture previously explained. The 2D router has four neighbouring ports to each side of the router and one local port to the network interface for the processor connection. We extended the 3D architecture implementation in this chapter by including processor architecture which allows us to investigate heterogeneous 3D architecture of complete MPSoC design because there exist both memory and logic structure. Figure 74 shows the interconnection structure between processor, NIU and 2D router for a complete tile block.

6.3.3 GALS Implementation

The GALS architecture is appealing from the power perspective where power reduction can be achieved due to the clock gating implementation whereas from performance perspective, it does not directly offers improvement which is depending on the implementation-specific techniques. A number of methods exists for interfacing different clock domains in the GALS architecture such as plausible clocking, FIFO-based and boundary synchronization as explained in details in [176]. One of the primary concerns of the GALS implementation is the data synchronization between different clock domains. Although FIFO-based GALS style suffers from the additional latency of the FIFO block, careful design and using large FIFO buffers can inherently hide much of the performance penalty [186] at the expense of more area overhead.

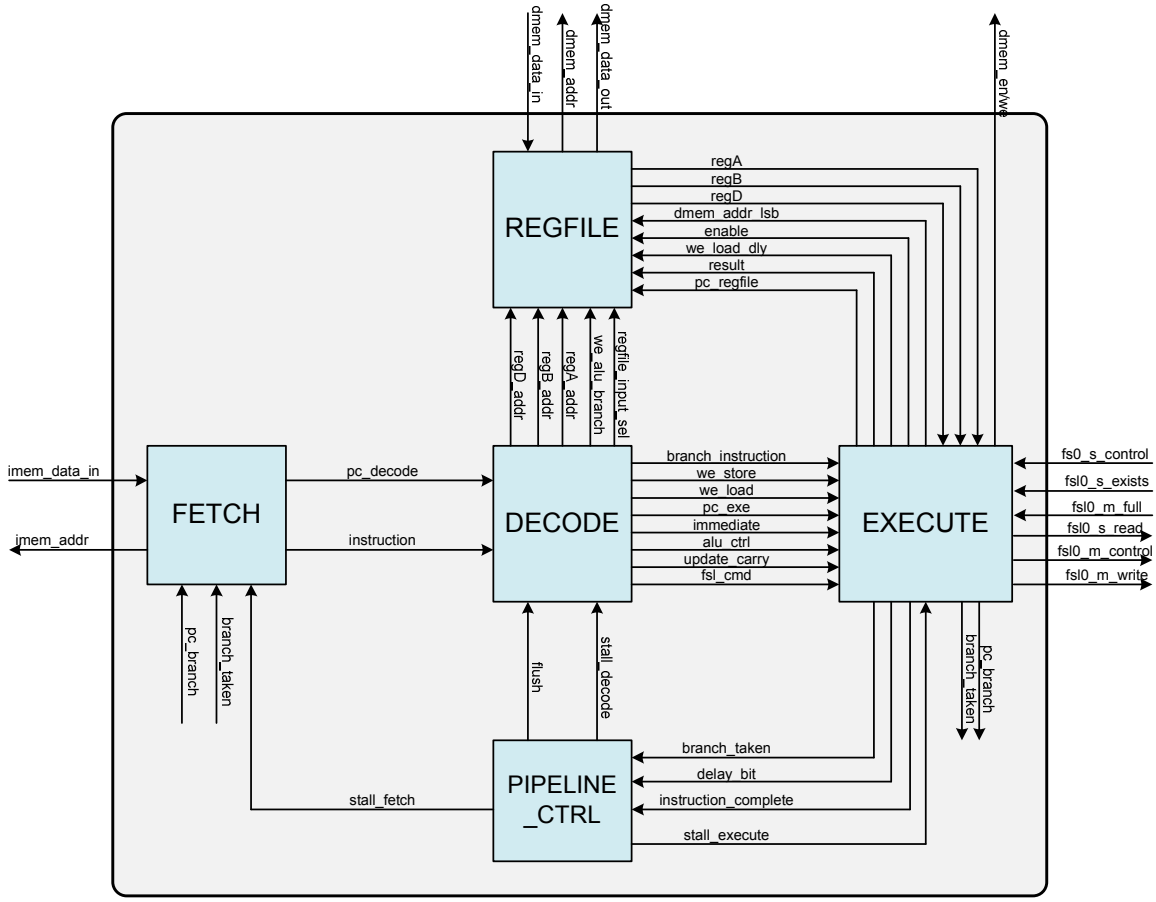


Figure 73: Openfire processor internal signals connection

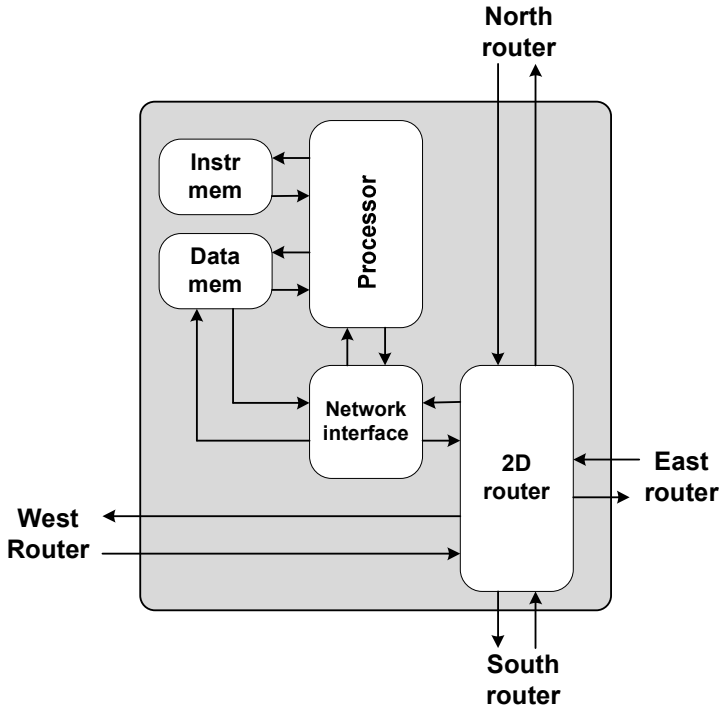


Figure 74: Interconnection structure for a complete tile block

The GALS style implementation in this architecture is depicted in Figure 75 which is based on a dual clock FIFO structure for handling clock domain crossing. We use a four-words depth for the FIFO block built-in within a network interface for transferring data from the processor through its FSL master and slave bus operating at 100 MHz to the NoC operating at 333 MHz. For processor to NoC communication, data from FSL bus is first written to the dual clock FIFO before being packetized to be sent to the router for transportation. In contrast, for NoC to processor communication, the packets arrive from router is first depacketized before being written to the dual clock FIFO.

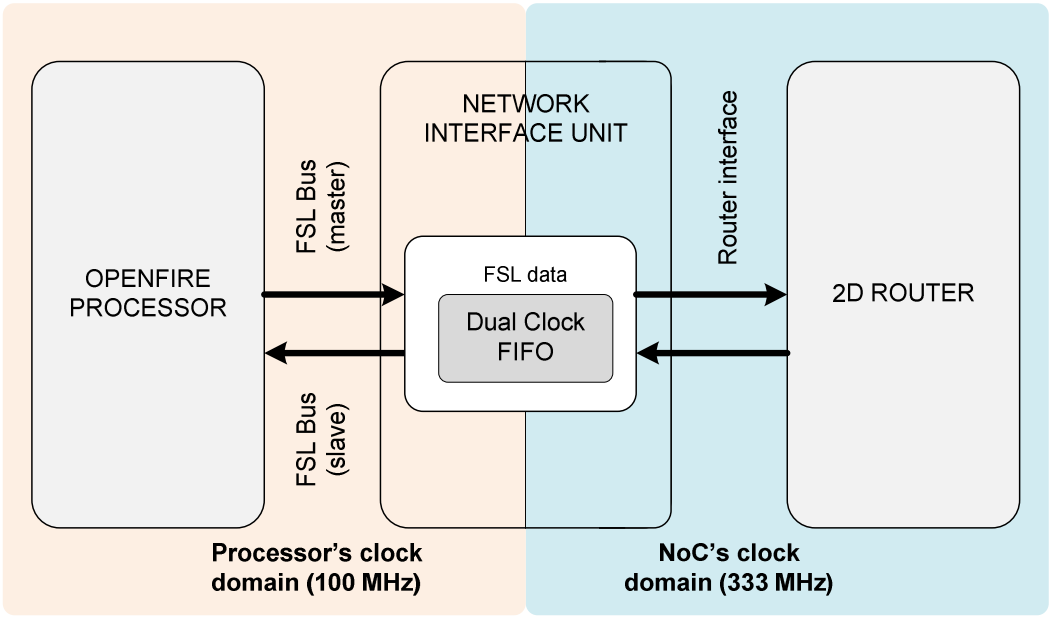


Figure 75: GALS implementation style using a dual clock FIFO architecture

6.3.4 Baseline 2D MPSoC Architecture

The 2D NoC-based multiprocessor architecture is shown in Figure 76 as a baseline design for comparison purposes with the heterogeneous 3D MPSoC stacking architecture. The synthesized area using 130 nm technology for each component is shown in Table 21 indicating that the tile area is dominated by the memory macros which is about 56% of the total tile area. We have implemented 16 processors with 4 KB data memory (dual port) and 4 KB instruction memory (single port) for each processor and using 2D Mesh NoC for the inter-processor communication based on the router and network interface explained in previously which consuming about 22 mm^2 silicon area using all metal layers available (up to metal 6).

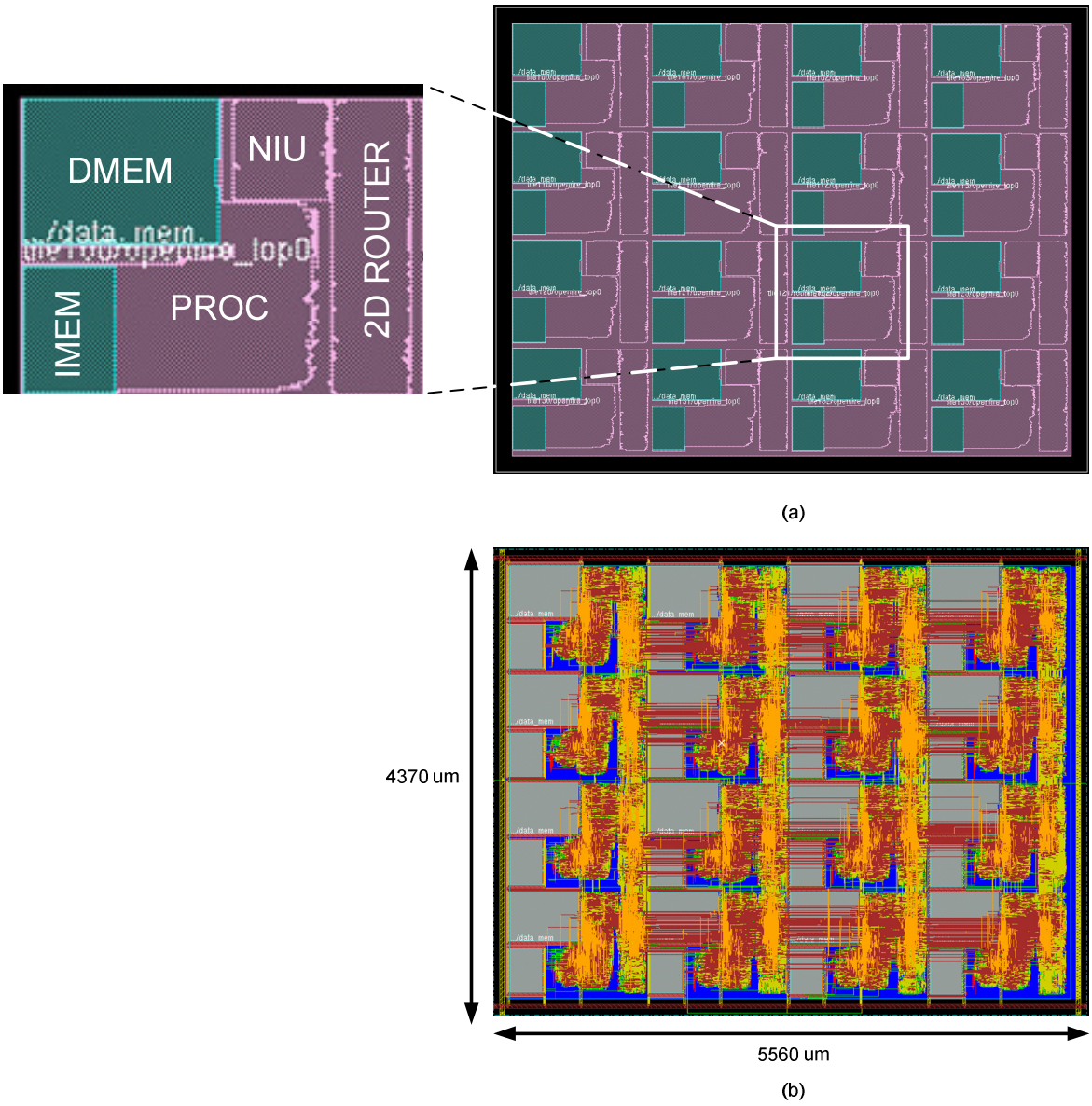


Figure 76: Baseline 2D MPSoC architecture (a) amoeba view (b) routed layout

Table 21: Synthesize area for each block in a tile

Component	Area (um^2)	Percentage (%)
Openfire CPU	161,035	18
Instruction memory (4 KB)	156,437	17
Data memory (4 KB)	352,550	39
NIU	63,100	7
2D Router	151,069	17
Total area per tile	884,191	100

6.4 Heterogeneous Stacking of 3D NoC-Based MPSoC Architecture

6.4.1 Partitioning Technique

For the heterogenous stacking, we divided the 2D design into a tile of processor and another tile for NoC architecture as shown in Figure 77. The floorplan and routed layout is shown Figure 78 and Figure 79 for bottom and top tier respectively. We use Tezzaron two-tier technology for this experiment and also the same design flow explained in chapter 3. The processor with its data memory is placed in the bottom tier while the NoC with the instruction memory is placed in the top tier. The vertical connection is made of signals from network interface in the NoC to the processor and to the data memory and also from the processor to the instruction memory. Therefore, first we set the location of the microbumps in the bottom tier around processors and data memory, then we floorplan the top tier for the NoC architecture by placing the network interface under the microbumps locations created from the bottom tier to be as close as possible. Stacking method proposed in [144] is not realistic because real routers have relatively small area compared with the processor or any other IP cores as fabricated in [187] and [188] which will creates large empty silicon area and therefore we decide to modify the floorplan by moving the instruction memory block to the top tier to be placed with the NoC architecture.

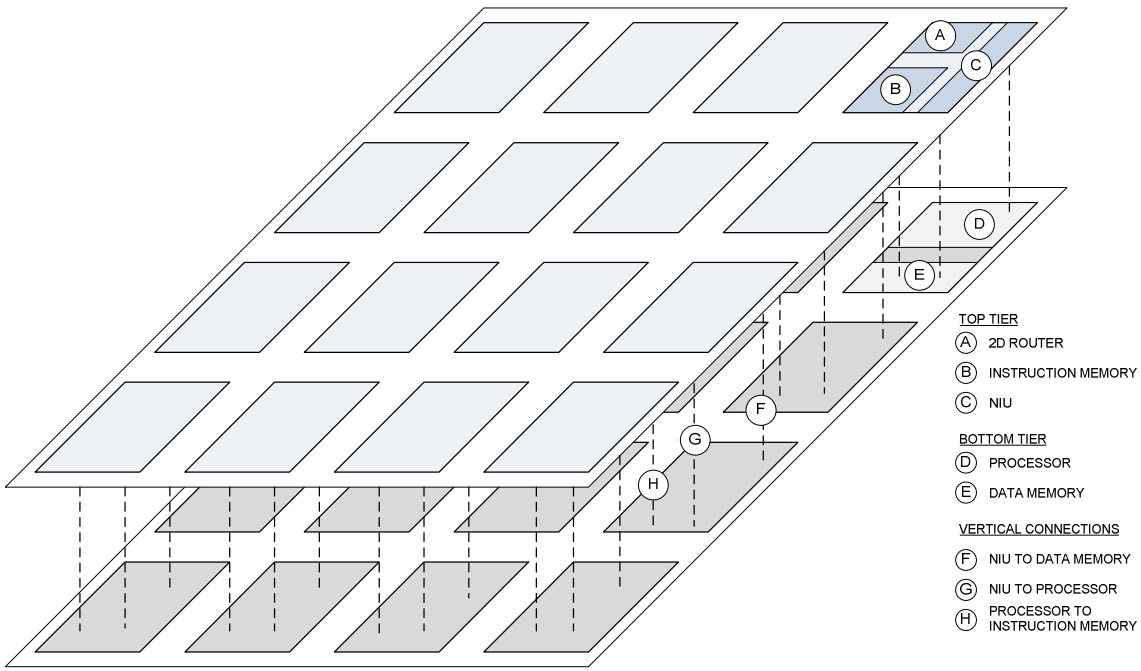
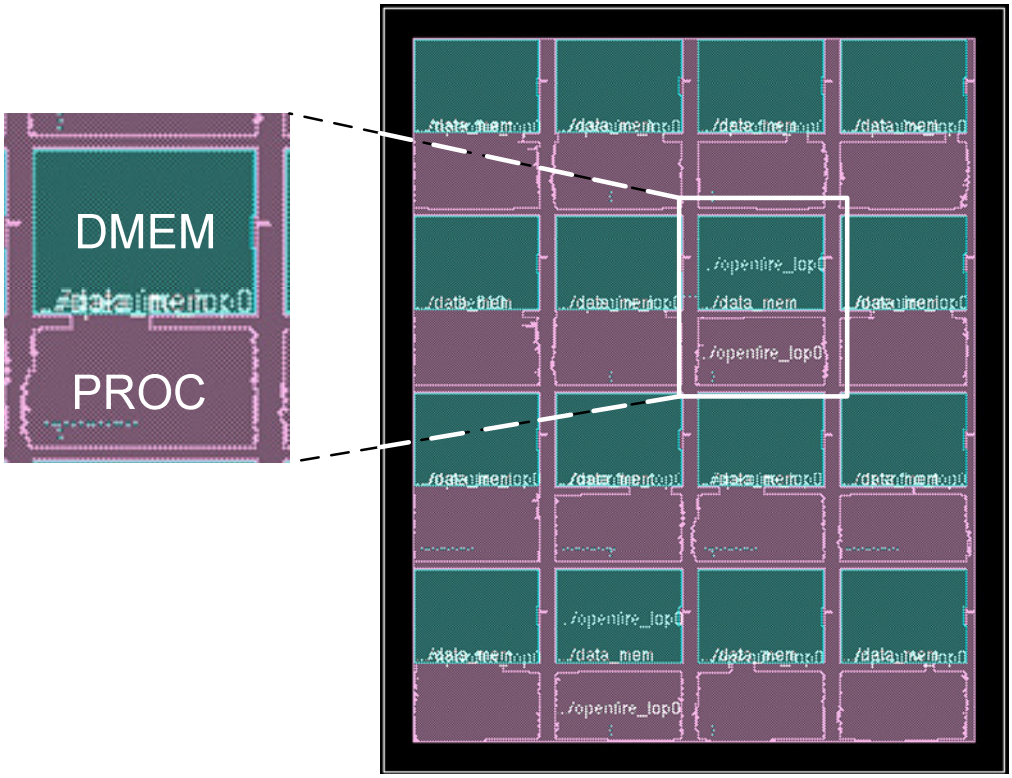


Figure 77: Heterogeneous 3D MPSoC stacking

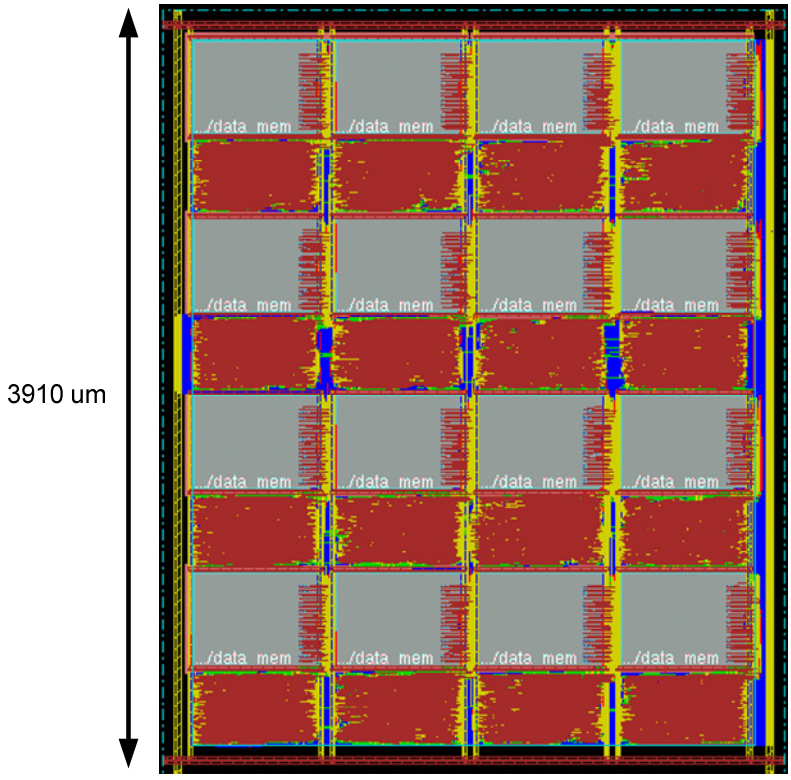
One of the novel features in this study is that we employ GALS in the 3D architecture wherein the NoC and the processor operate in different clock domains since the processor is quite slow compared with the speed of NoC. To the best of our knowledge, this work is the first to conduct physical design implementation analysis of 3D GALS for multiprocessor architecture. The GALS clocking style avoids global clock tree structure which essentially reduces power consumption since clock tree has prominent portion of the total power consumption of a system. A part from that, this implementation style also enables Dynamic Power Management (DPM) and Dynamic Voltage and Frequency Scaling (DVFS) [175] methods for balancing power consumption and performance at real time and also allows efficient thermal management specifically for 3D architecture having higher temperature effect. Based on the GALS architecture, each tier can be run at different frequencies where the NoC at the top layer is clocked at 3 ns while the processor at the bottom layer is clocked at 10 ns period. This type of floorplan provides easier thermal management technique by placing the hot layer clocked at higher frequency close to the heatsink enabling fast thermal transfer [170]. From the testing point of view, this floorplan also allows easier method for 3D architecture pre-bond testing of the NoC as well as processor architecture since they are located in separate layer.

6.5 Experimental Results

It can be seen from Table 22 that there is almost 50% reduction of core area for heterogeneous 3D stacking compared with the 2D architecture due to the partitioning of NoC architecture and instruction memory into another layer. The number of gates however is slightly increased over 2D architecture mainly because of separate optimization flow of both tiers during place and route step. Out of 188 vertical connections per tile (NIU to/from processor and data memeory), 70 connections are for the processor FSL connections whereas the rest of vertical connections are for the data and instruction memory connections. We can also see the slight increased of total wirelength in heterogeneous 3D stacking compared with the 2D architecture due to separate 2D optimization process during place and route step. As shown in Table 22, the speed of the NoC is improved in 3D architecture.

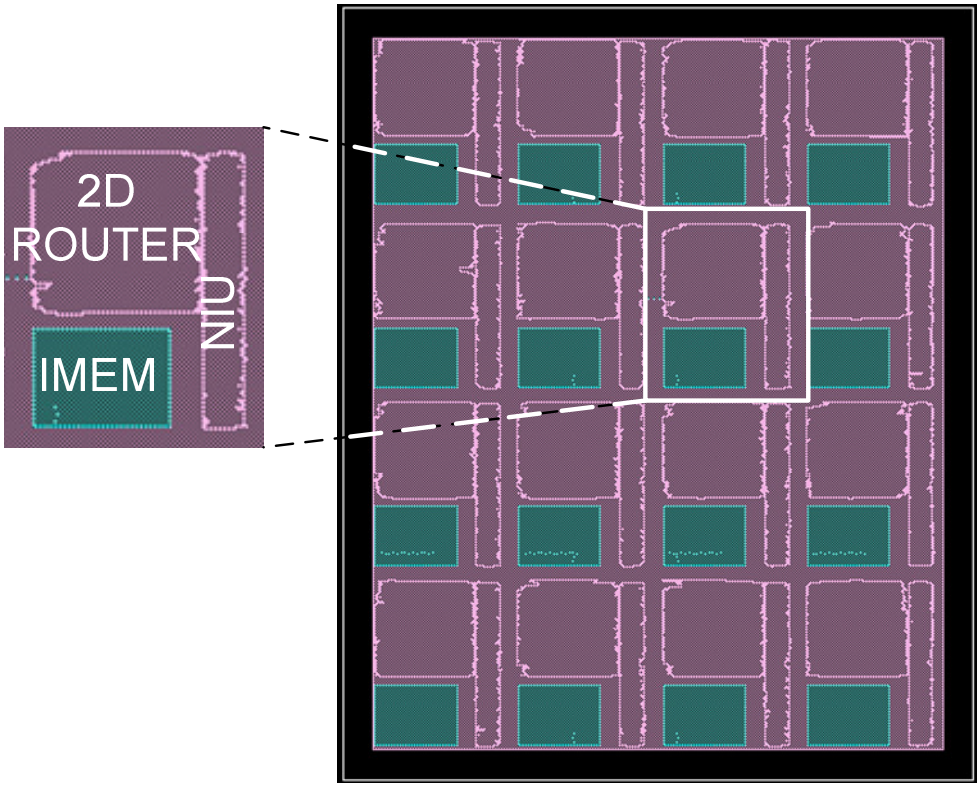


(a)

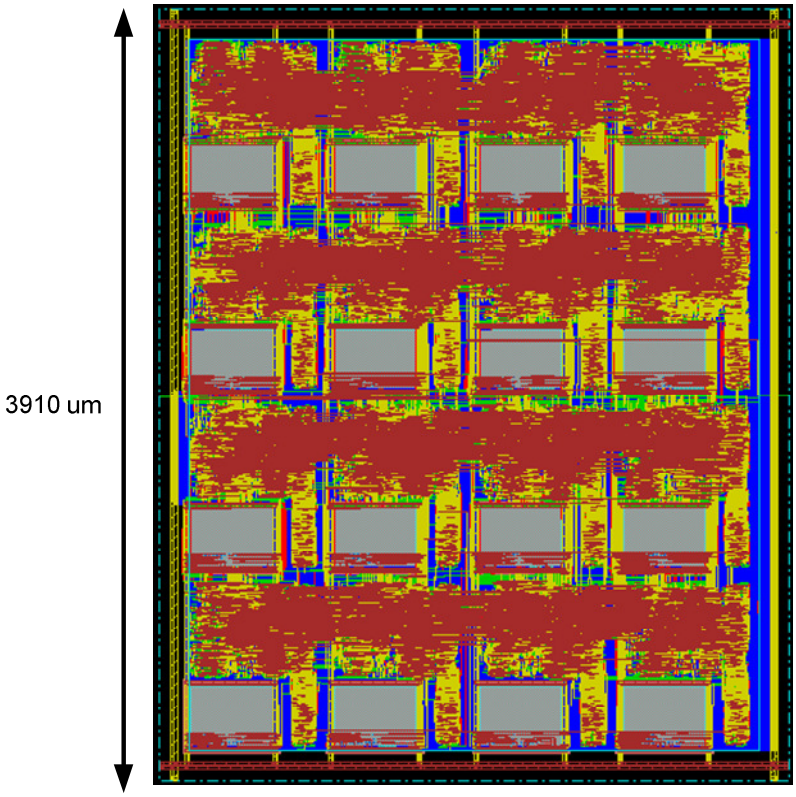


(b)

Figure 78: Bottom tier of heterogeneous 3D stacking (a) amoeba view (b) routed layout



(a)



(b)

Figure 79: Top tier of heterogeneous 3D stacking (a) amoeba view (b) routed layout

Table 22: Performance comparison for 2D and 3D heterogeneous stacking

Parameters	2D architecture	3D heterogeneous stacking
Core area (mm^2)	21.4	10.4
Number of gates (million)	2.70	2.73
Number of total microbumps	-	3011
Number of microbumps per tile	-	188
Microbumps for IMEM per tile	-	42
Microbumps for DMEM per tile	-	76
Microbumps for FSL per tile	-	70
Total wirelength (m)	21.1	21.4
Critical path delay for NoC clock (ns)	3.51	3.19
Critical path delay for processor clock (ns)	9.92	10.09
Power Consumption (W)	1.38	1.48

The performance comparison between 2D and 3D design is shown in Table 22 and Figure 80 where it clearly shows that heterogeneous 3D stacking improves slightly in the NoC speed. Performance increased in the NoC speed is partially because of the area reduction which contributes to wirelength reduction for the critical path (from input to register path). In terms of power consumption, the marginally increased of 3D architecture power consumption over 2D architecture is due to the increased of logic gates in 3D architecture as well as its total wirelength as a result of separate place and route run for each tier.

Figure 81 shows the horizontal wirelength distribution of 2D MPSoC, bottom tier and top tier of heterogeneous 3D stacking where below 0.8 mm length, it can be seen that the number of wires for the heterogeneous 3D stacking is decreased but have more wires for wirelength between 0.8 mm and 0.9 mm. As we run separate place and route for each tier, therefore the tool will optimize each tier accordingly without considering the complete 3D architecture which could be the reason of this trend.

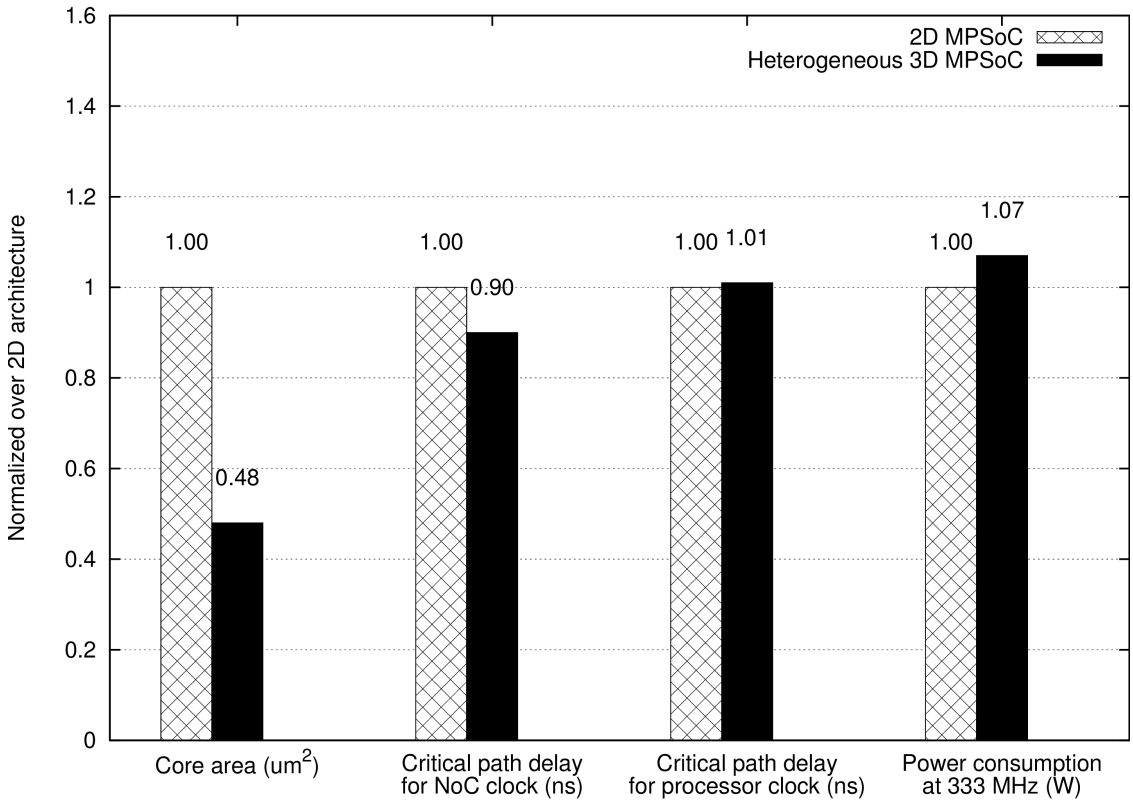


Figure 80: Performance comparison for 2D and heterogeneous 3D MPSoC architecture

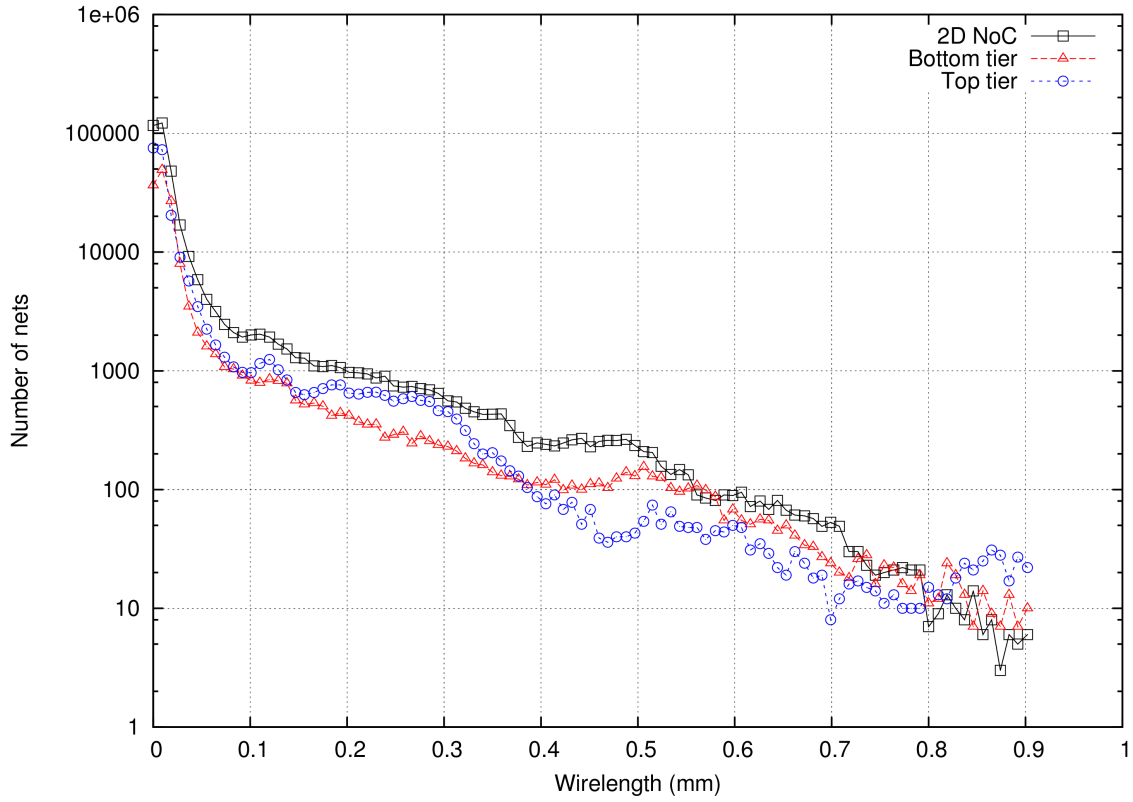


Figure 81: Horizontal wirelength distribution for 2D MPSoC and 3D MPSoC (bottom and top tier)

6.5.1 2D vs 3D Clock Tree Analysis

Clock tree synthesis for 3D architecture has been studied especially for synthesizing clock tree in many tiers targeting low skew as well as low power consumption. In [189], several clock tree topologies have been analyzed based on the fabricated three-tier 3D chip using MIT LL technology. Measured data from the fabricated chip suggesting that the H-tree structure gives the lowest skew but highest power consumption compared with the other clock tree structures. Several clock tree schemes have also been proposed considering various objectives such as timing yield, fault tolerant, TSVs blockage problem, testability and process variation between dies and within a die [190] [191] [192] [193] [194] .

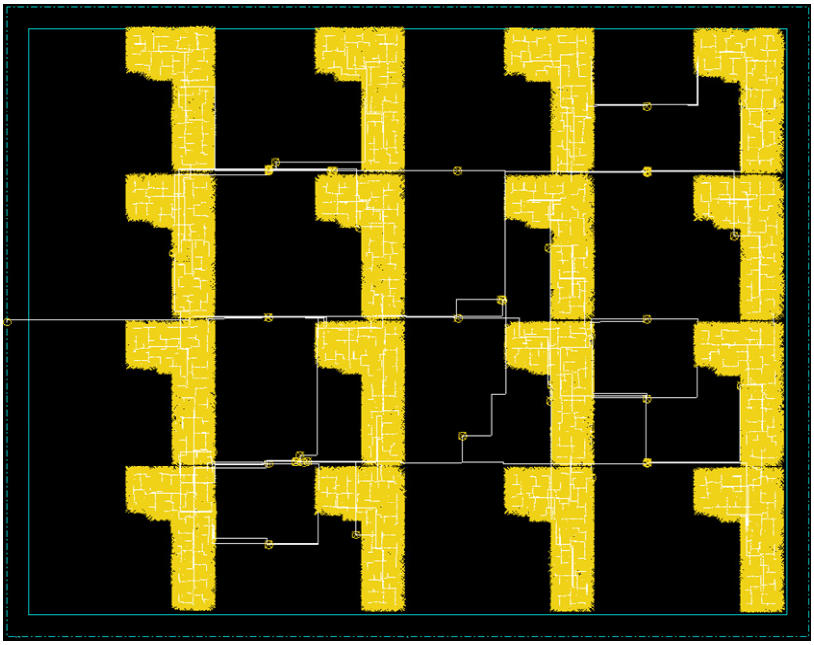
Several physical design implementations of 3D architecture has been reported previously conducting performance analysis based on layout-level netlist. However, there is no details discussion regarding the implications of the generated clock tree structure using 2D CTS tools to the overall 3D clock tree structure. Even though there are some works used 2D tool to generate the clock tree [195] [196], nonetheless they did not measure the impact of the method to the 3D timing performance which is the aim of this particular discussion. In this section, comparison of clock tree structure between the baseline 2D architecture and heterogeneous 3D stacking is carried out to have better insight as well as to highlight issues related to the 3D clock tree structure.

One of the benefits of deploying GALS architecture is that we are able to control the rising value of clock skew in the fully synchronous implementation especially for advanced technologies where very dense clock tree structure is created due to the higher registers density. The higher level of clock tree structure increases the clock skew value as well as more sensitive to the on-chip variation (OCV) [193]. In GALS architecture, as the clock skew constraints is limited only to its block boundaries thereby open up design spaces for performance enhancement as well as less hardware requirement since the complexity of the clock distribution is reduced.

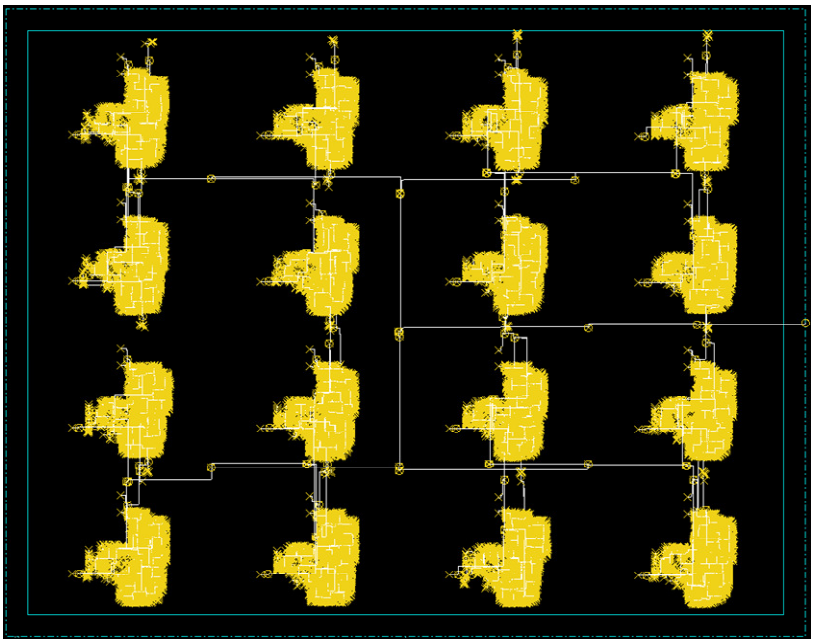
The clock tree structure for both 2D MPSoC and heterogeneous 3D MPSoC stacking are shown in Figure 82 and Figure 83 for both NoC and processor clocks in bottom and top tier respectively. The clock tree synthesis for both architectures is done automatically by the CTS Engine in SoC Encounter where the clock specification file is generated based on the supplied timing constraints. A microbump per clock signal has been placed at the centre of the top tier to enable balance distribution between both tiers from the clock source that coming from the top tier. As shown in the figures, CTS Engine synthesized the clock tree with H-tree topology at the first 3 or 4 levels. Table

23 presents the clock tree synthesis structure between 2D and 3D design where it is clearly shown that the clock tree structure of 3D design (combine both bottom and top tiers clock tree structure) for processor clock and NoC clock have less number of clock tree level compared with the 2D design. For the number of sinks and number of buffers, the difference between 2D and 3D design is not very significant for both processor and NoC clock which is indicating that 3D design does symplify the clock tree structure through reducing the number of clock tree level for the same number of sinks and buffers. Another point is that generating clock tree synthesis in 3D design using 2D physical design tool does not have differ substantially whether the clock tree structure is exist only in a single tier of the 3D design or exist in both tiers.

Reduction of the number of clock tree level could potentially improve power consumption where clock network has substantial portion of total power consumption in a chip especially in advanced technology [197]. However, as shown in Table 23, the clock skew of processor clock in 3D architecture is larger than in 2D design whereas NoC clock the opposite trend. The possible reason for the large skew of processor clock in 3D architecture is because the processor clock tree for both tiers has been generated and optimized separately during place and route step which although the optimization process is able to reduce the number of clock tree level, however the tool does not able to minimize the clock skew because it does not see the complete 3D architecture during the optimization process.

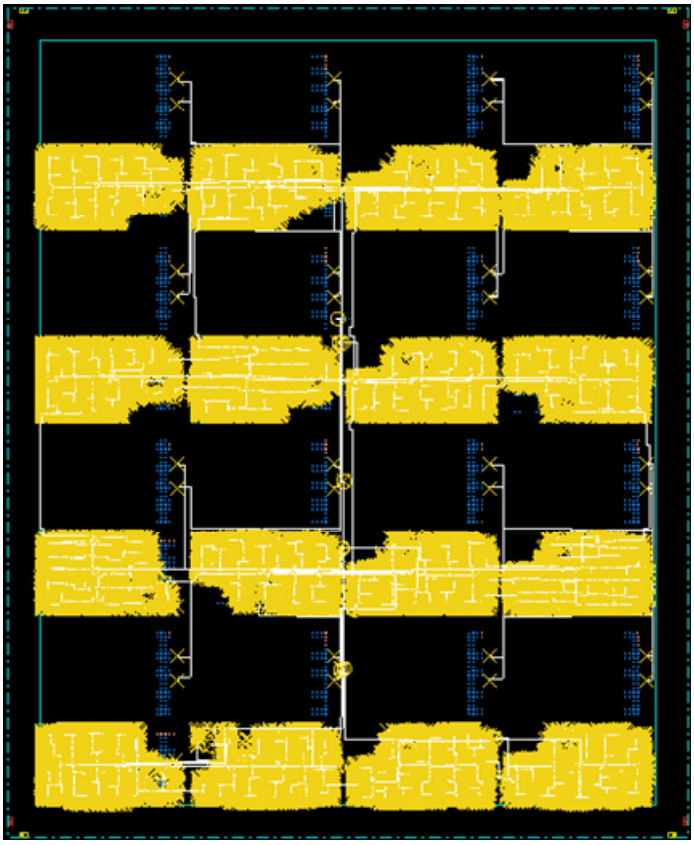


(a)

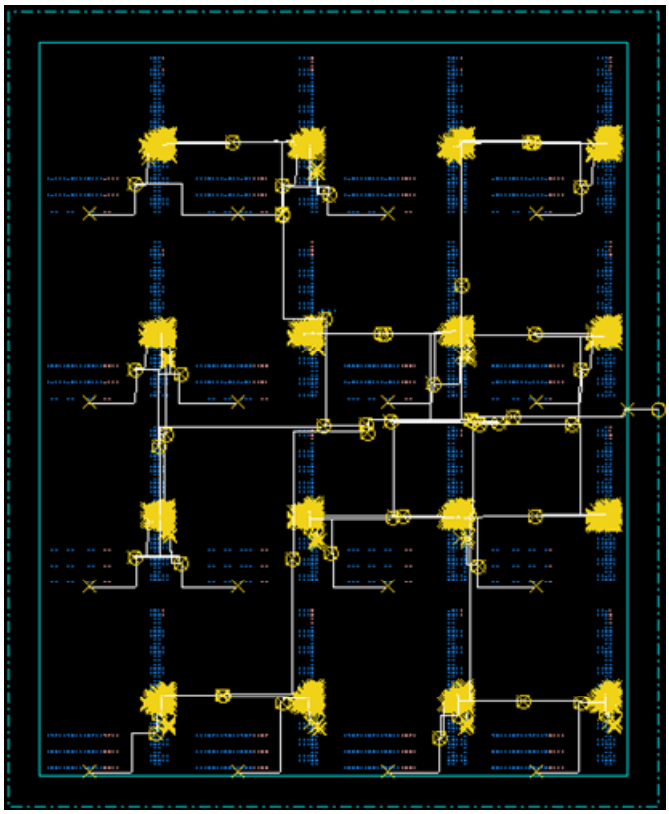


(b)

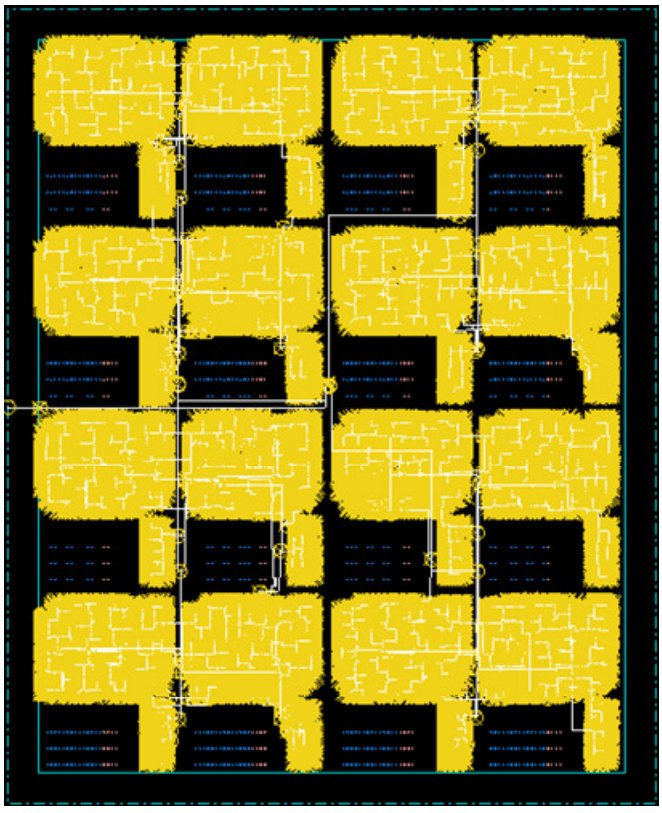
Figure 82: Clock tree structure of 2D MPSoC architecture (a) NoC clock (b) processor clock



(a)



(b)



(c)

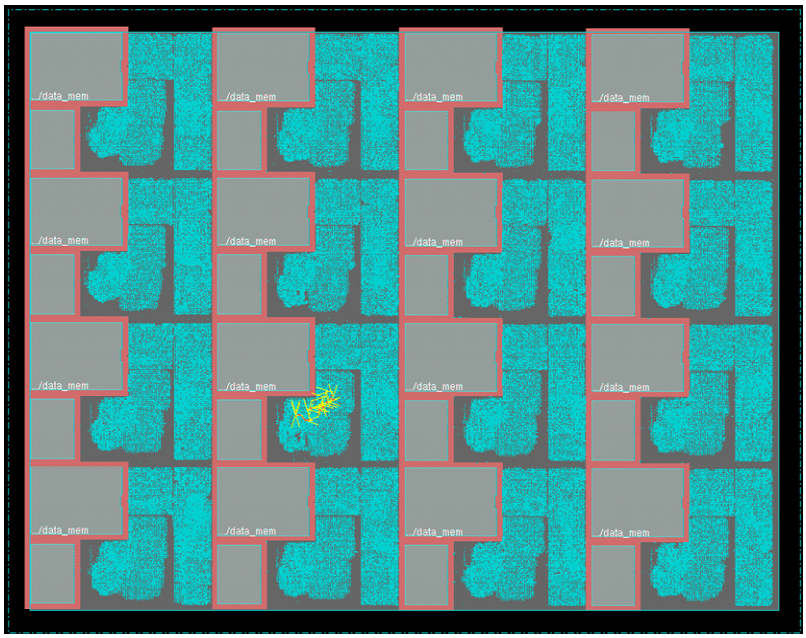
Figure 83: Clock tree structure for heterogeneous 3D MPSoC stacking (a) processor clock of bottom tier (b) processor clock of top tier (c) NoC clock of top tier

Table 23: Clock tree structure properties for 2D and 3D designs

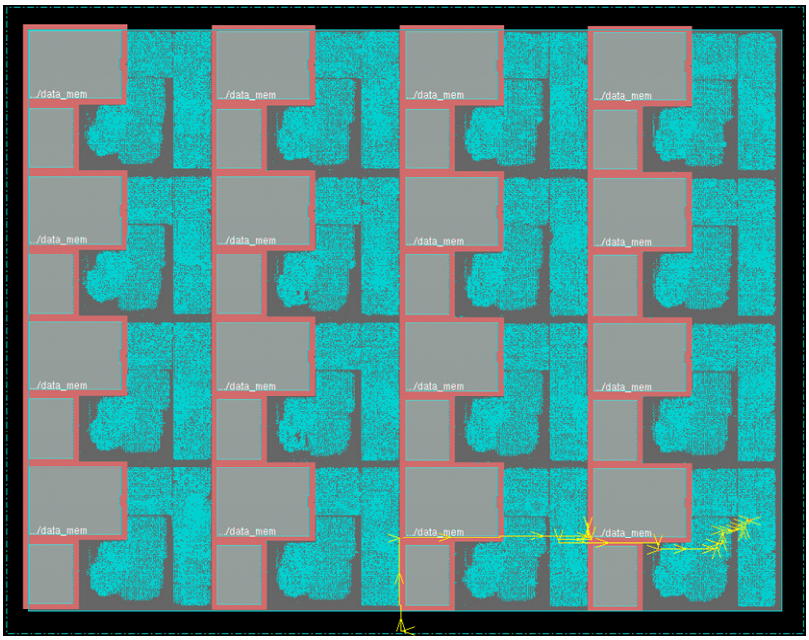
Parameters	2D		3D (bottom tier)		3D (top tier)	
	Processor clock	NoC clock	Processor clock	NoC clock	Processor clock	NoC clock
Level	17	10	7	-	6	8
Number of buffers	944	1580	879	-	72	1599
Number of sinks	40928	72832	38640	-	2288	72832
Skew (ns)	0.40	0.43	Processor clock skew = 0.76 NoC clock skew = 0.07			

6.5.2 2D vs 3D Critical Path Analysis

By analyzing the critical paths of both architectures, we will get better insight regarding how we can properly apply the 3D design constraints to get higher performance improvement. Figure 84 and Figure 85 show the critical paths for processor clock and NoC clock for both 2D and 3D architectures respectively. In this 3D architecture, the critical paths for both clocks reside in the blocks (inside the router for NoC clock and inside Openfire block for processor clock), meaning that the partitioning methodology that divides the 2D architecture at block-level partitioning in 3D architecture does not affect the original critical paths as in 2D architecture. Therefore, the impact of designing 3D architectures based on heterogeneous stacking at block-level partitioning does not have prominent impact to 3D performance due to the fact that conventional 2D physical design tools can be used to accurately perform design optimization without having to see the complete 3D architecture. Unlike in homogeneous 3D stacking where the performance of the architecture is determined by the 3D critical paths, heterogeneous 3D stacking architecture at block-level partitioning has performance affected by the 2D critical path within the block implementation. The FIFO architecture is used for handling clock domain crossing and is placed inside the NIU in the top tier where the processor signals are coming from the bottom tier through vertical connections. Hence it provides easier design verification as well as easier pre-bond testing due to separate architecture with separate clock frequency.

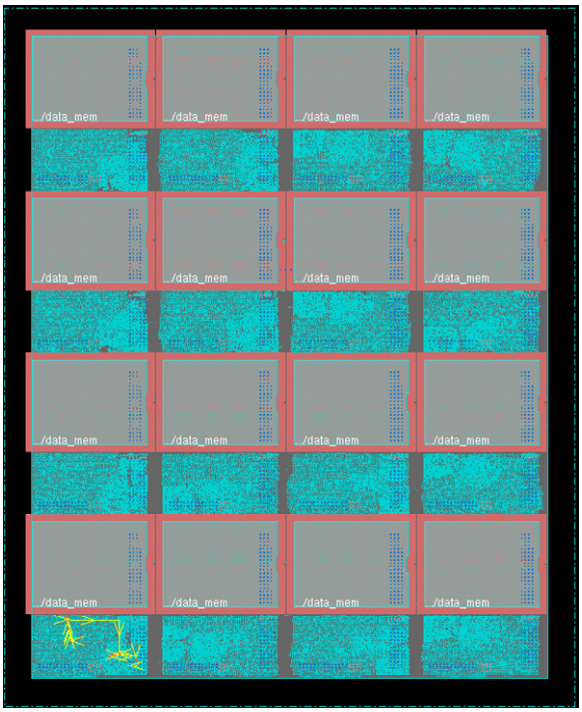


(a)

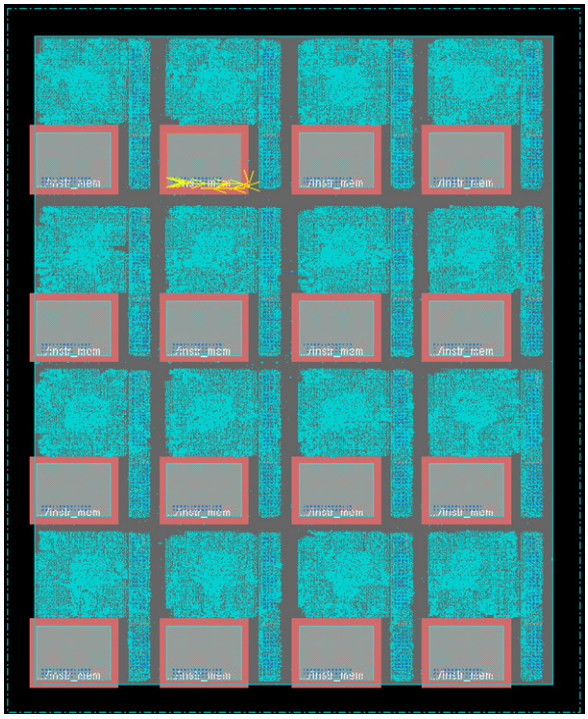


(b)

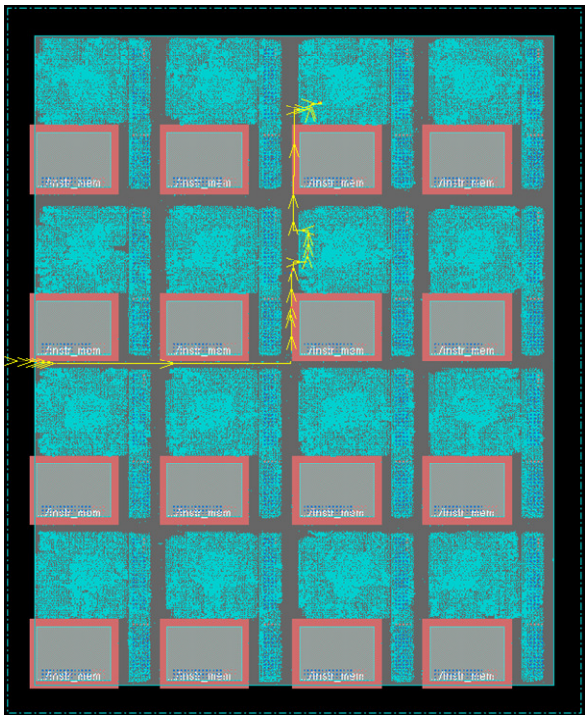
Figure 84: Critical path for 2D MPSoC (a) processor clock (b) NoC clock


../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem
../data_mem

(a)


../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem
../instr_mem

(b)



(c)

Figure 85: Critical paths of each tier separately in SoC Encounter for the heterogeneous 3D MPSoC (a) processor clock in bottom tier (b) processor clock in top tier (c) NoC clock in top tier

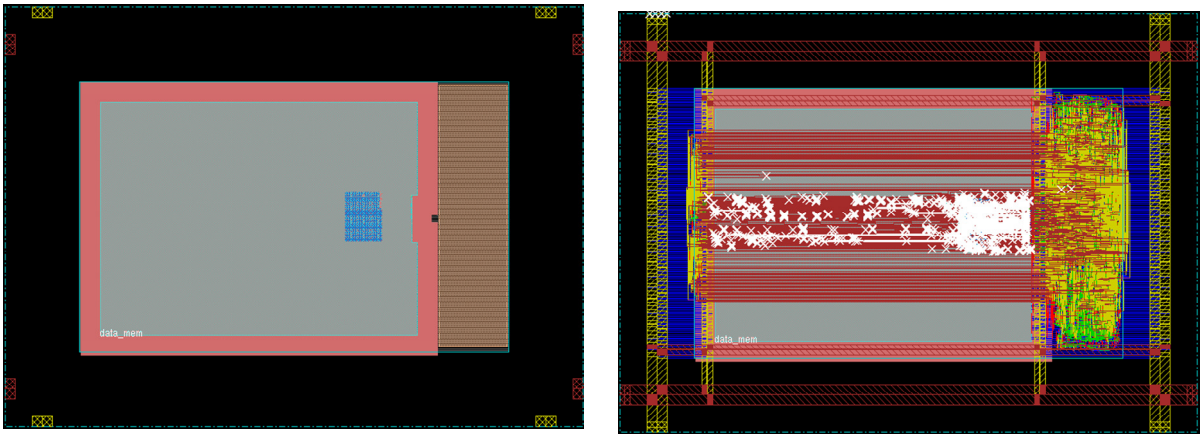
6.5.3 Impact of Microbumps Pitch

To further analyze the trade-off in physical design of 3D IC design using 2D EDA tool, we carried out an experiment to measure the implication of microbumps pitch for the vertical interconnection to the performance. The fundamental reason is that although microbumps does not create routing block and does not have large keep-out-zone as TSV because it uses top metal layers and the routing is done until one layer below top metal layer, doing place and route using 2D tools can contribute to the routing congestion in the area near to the microbumps when pitch is very small hence limiting the optimization process due to the dense routing space between microbumps structure. This is not problematic when designing with 3D tools because the tools allow optimization of the cells between tiers together with its vertical assignment.

Figure 86 and Figure 87 show the physical design implementations for different microbumps pitches where we place and route the design and perform timing analysis to compare its effect to the 3D timing characteristics. We only use Openfire processor for this experiment to keep it simple as the purpose it to see the impact of microbumps pitch for the vertical signal assignments. The processor is divided into two tiers where the data memory and execute block is placed in bottom tier

while the rest are placed in the top tier where critical path is located between execute block and decode block (3D critical path) that have been analyzed from the 2D implementation. Arranging microbumps arrays with large pitch such as 20 µm over logic and memory does not produce any problem as the place and route is able to route the design and do the optimization. However, for smaller pitch such as 5 µm, placing microbumps array over memory macro block create many DRC violations due to the insufficient metal connection to route the signals to the microbumps. When placing this microbumps array with small pitch over the logic block, there is no DRC violation reported by the place and route tool and optimization is performed successfully. This issue shows that microbumps structure for face-to-face 3D stacking provides benefits over using TSVs but impose some architectural constraints especially for very high density vertical interconnections with many or large memory blocks.

Measuring the timing performance for microbumps with 5 µm pitch and 20 µm pitch shows quite significant impact of microbumps pitch to the 3D timing performance in this design even though this design has relatively small area to represent a realistic design. We perform 3D timing analysis by feeding the RC parasitic files of both tiers generated from SoC Encounter tool to the Synopsys PrimeTime tool and analyzed the critical paths for both designs. Table 24 shows comparison of slack and clock skew between both implementations for a target clock period of 10 ns where it clearly shows that design with larger pitch has more slack but does not affect the clock skew when compared with the smaller pitch. Even though the 3D timing is also affected by the location of the microbumps due to the horizontal wire length before reaching the microbumps, the microbumps pitch contribute insignificantly to the 3D timing paths because the length difference is relatively small compared with the implementation with larger microbumps pitch. Therefore, smaller microbumps pitch provides higher vertical interconnection density but still needs to be optimized with the target architecture considering its location for the signal assignments.



(a) (b)

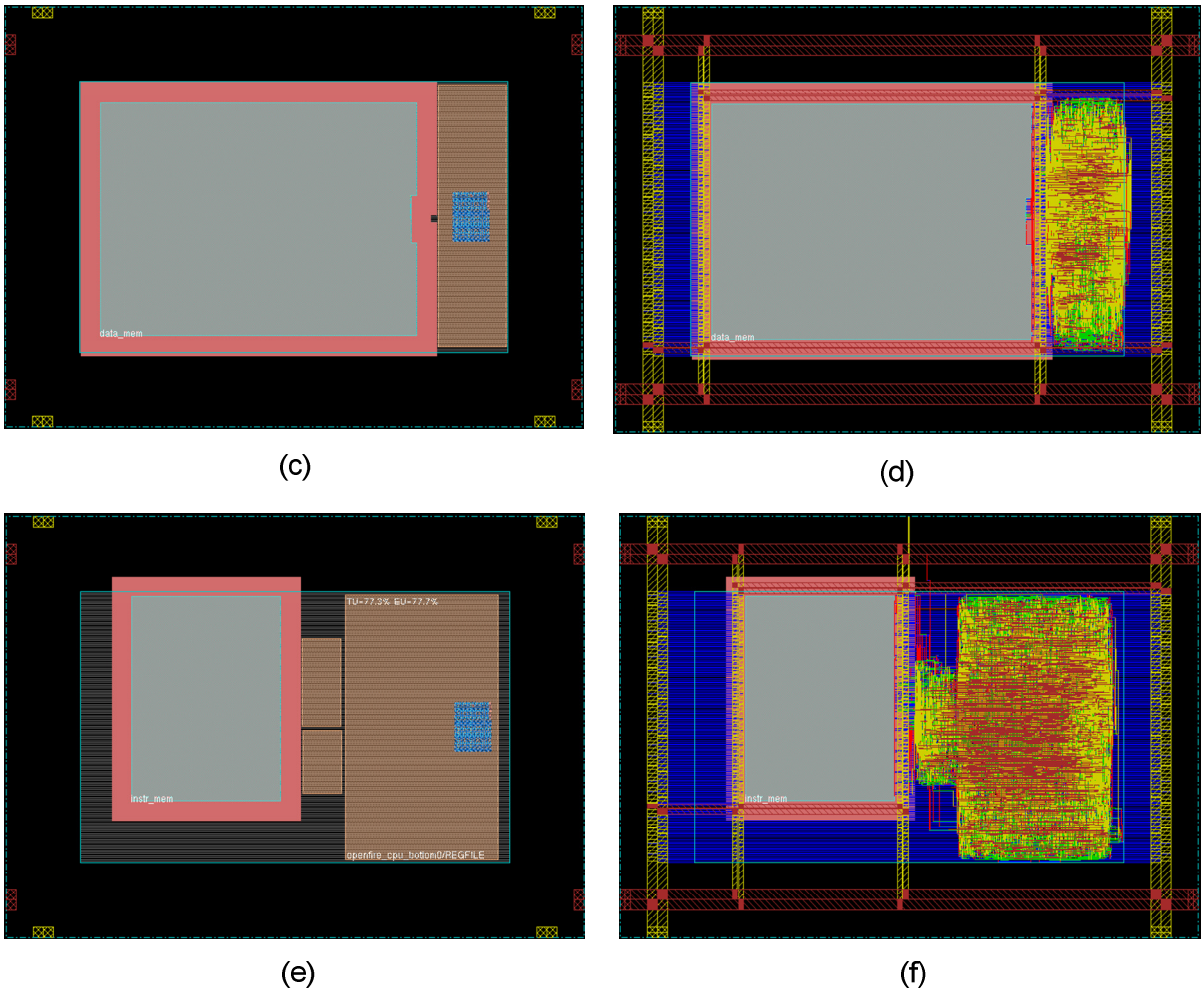


Figure 86: Openfire 3D architecture with 5 µm microbump pitch (a) floorplan of bottom tier with microbumps array on top of memory block (b) routed layout of bottom tier with many DRC violation (c) floorplan of bottom tier with microbumps array on top of processor logic (d) routed layout of bottom tier (e) floorplan of top tier (f) routed layout of top tier

Another point to note is that microbumps also create power structure blockage due to the fact that metal 5 is used for signal connections to the microbumps (metal 6) and is also used for the power structure where upper most metal layers are the preferred layers due to their smaller resistance in order to reduce IR drop. Power ground network structures such as around macro blocks and vertical or horizontal stripes restrict the microbumps array location which has severe effect if there is a high number of a vertical connection. Hence, power ground network and microbumps locations must be co-optimized during floorplanning stage to achieve target performance requirement as well to ensure sufficient power delivery.

Additionally, despite the advantages of microbumps architecture in physical design as mentioned earlier compared with TSV, it also induce stress to the transistor like the TSV as reported recently by [198]. Therefore, in addition to the microbumps pitch, this stress effect should also be analyzed to evaluate its impact to the 3D timing as have been done previously for the TSV [111] to be able to identify the issues of employing this technology in 3D architecture.

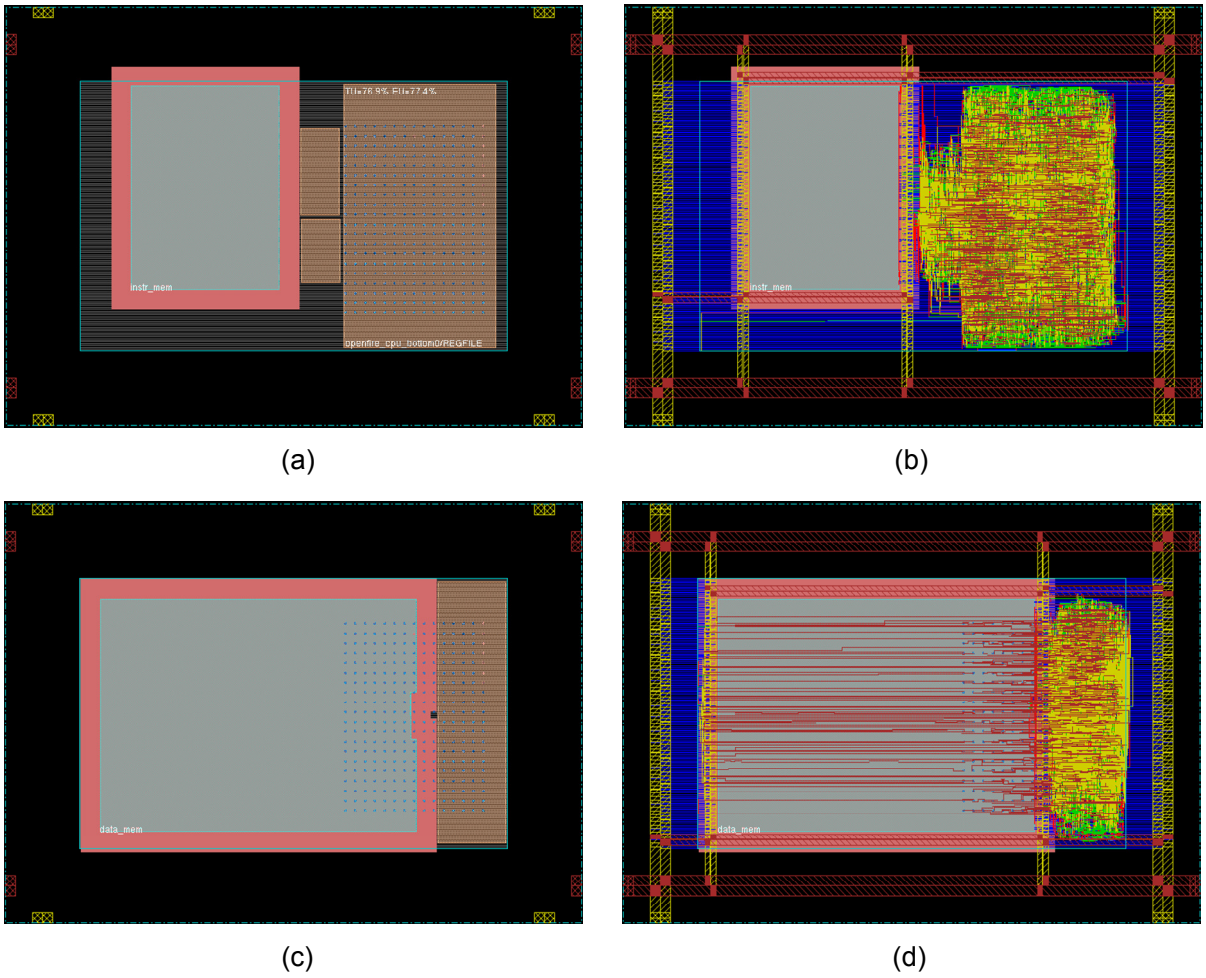


Figure 87: Openfire 3D architecture with 20 µm microbumps pitch (a) floorplan of top tier (b) routed layout of top tier (c) floorplan of bottom tier (d) routed layout of bottom tier

Table 24: Timing performance of different microbumps pitches (target clock period of 10 ns)

Parameters	5 µm pitch	20 µm pitch
Timing slack (ns)	0.08	0.72
Clock skew (ns)	0.52	0.58

6.5.4 Implications of 3D IC Design using 2D EDA Tools

One of the primary limitation of using 2D EDA tools for designing and implementing 3D IC architecture is the lack of design exploration support. To be able to gain as much performance as possible from the 3D technology, the need for design exploration is utmost important to evaluate different implementation trade-offs for a specific target hardware or application before proceeding with complete design implementation flow. Specific to the heterogeneous 3D stacking at block-level partitioning, as long as the critical paths reside inside the block architecture thereby using 2D EDA tools seem to be sufficient enough to be able to design as well as doing optimization due to the fact that the tools does not require to see the complete 3D architecture.

Although manual design exploration can be carried out using the 2D tools as have been done for the microbumps pitch exploration, it is not sufficiently accurate analysis because generally the parameters in 3D architecture is interrelated to each other and thus must be done in a complete 3D flow integration using 3D tools for more accurate exploration analysis. For example, exploring microbumps pitches for the vertical signals assignments must be carried out simultaneously with the microbumps location explorations to be able to accurately estimate its performance impact to the 3D architecture.

Conclusion

In this chapter, we have discussed heterogeneous 3D stacking of NoC-based MPSoC architecture. We explored other feasible 3D architecture implementation of MPSoC architecture compared with the homogeneous 3D stacking architecture discussed in previous chapter to analyze its performance as well as to have more understanding with regards to the architectural design trade-offs. The GALS style implementation provides benefits due to separate clock domains between communication and computation architecture which could be the main interest for employing it in the 3D architecture. One of the important points in designing 3D architecture for heterogeneous 3D stacking architecture with block-level partitioning, 2D EDA tools can be used as in a normal flow in 2D design by carefully partitioning the design to have 2D critical paths located within a tier and thus does not need 3D optimization process. Furthermore, although face-to-face stacking using microbumps does not have problems such as routing blockage and large keep-out-zone area (as opposed to TSV), it does imposes physical implementation constraints where there exists routing congestion leading to many DRC violations for very small microbumps pitch. This is especially critical for complex memory dominated 3D architectures which need high density inter-tier connections.

CHAPTER 7
DESIGN SPACE EXPLORATION OF 2D EDA TOOL IMPACT ON THE 3D MPSOC ARCHITECTURE

7.1 Introduction

Design space exploration is one of the important things to be concerned helping designers to evaluate different possible design implementations before any design is implemented in real hardware. Design space exploration refers to an activity of exploring design alternatives which commonly implemented using high level approach with the support of accurate modeling tool. The main reason of choosing high level approach is to reduce tool run time as there is a vast design space to be explored and tool run time it crucial where it will limits the exploration activities if it require very long time for each exploration.

Compared with 2D architecture, design space exploration of 3D architecture is more complex due to the much higher design space to be explored with the new structure which is vertical connection using microbumps or TSV. Fully automatic 3D design space exploration requires an accurate model of physical structure of the 3D architecture in order to have more useful evaluation for the target hardware implementation. One of the primary issues in doing 3D architecture design is the use of 2D EDA tool which is not an appropriate method since the tools is mainly dedicated and optimized for 2D architecture design. Therefore, using the tools to design and optimize 3D architecture could have strong or no impact at all to the 3D architecture performance. The aim of this study is to evaluate how the 2D EDA tools affect the 3D architecture performance when varying the tools options such timing slack, power consumption, gate count and total wirelength.

In this chapter, we present a design space exploration 2D EDA tool impact on the 3D design and implementation to highlight issues and challenges pose from this design methodology. Due to the unavailability of 3D design tools capable of doing 3D synthesis, 3D placement, 3D CTS (clock tree synthesis) and 3D routing, designing using 2D EDA tools especially for place and route is the common solution. The aim of this study is to analyze how 2D EDA tools are affecting the overall 3D architecture performance, specifically timing performance and power consumption, pointing out important points related to the design and optimization of 3D architecture.

7.2 Related Works

A number of works have been reported with regards to the design space exploration of 3D architecture. System level design space exploration for 3D architecture is proposed by [199] enabling exploration of different stackings and partitioning schemes and their affect on the performance, power and temperature. The proposed design space exploration is supported by several high level estimation tools such as GEMS TLM (transaction level modeling) for multicor performance analysis, ORION for NoC power analysis and WireX for thermal analysis. Another design space exploration for 3D stacked architecture is presented in [200] [201] focusing different 3D packaging solutions with logic and memory integration. The proposed flow is demonstrated with the video enconding applications with the support of commercial tools such as CoWare for high level synthesis and Javelinj360 for physical design prototyping.

In [202] [203], architectural-level exploration framework for 3D SoC architecture has been proposed that is tuning to optimize power/energy targeted for embedded systems. Commercial 2D EDA tools as well as a novel 3DPart tool for partitioning functional blocks into several tiers are integrated in the exploration framework for 3D system prototyping to evaluate its performance. A multi-criteria decision aid (MCDA)-based design space exploration has been proposed by [204] to deal with the growing complexity in 3D architecture due to the huge solution spaces. Establishing Pareto frontier in the multi-objective optimization exploration process is implemented by means of Non-dominated Sorting Genetic Algorithm (NSGA) before best decision is made using MCDA tool. Fast design space exploration using high level exploration framework has been presented by [205] to solve the problem of exploring huge design spaces and has been demonstrated using fairly complex MPSoC platform running AVC/H.264 video encoding applicaition. The high level exploration framework based on hierarchical mapping model developed using C++ with XML interface is validated for its accuracy with the low level framework that is based on extended version of transaction level WormSim NoC simulator and SoC Encounter physical design tool.

Differs from the previous reported works, this study perform a design space exploration of 2D EDA tool parameters on the 3D MPSoC architectures. Our previous work of the exploration is limited to several things such as exploration on a single tier, only explores placement options and did not use the 3D MPSoC with NoC architecture (using simple architecture which is Filter which requires short tool run time during the exploration) [206]. Therefore it did not provide accurate analysis on the impact of the 2D tool on 3D architecture performance especially for 3D MPSoC architecture performance which is the objective in this experiment. We have chosen different placement and

routing options in SoC Encounter and evaluate their impact on the timing slack, power consumption, gate count and wirelength of the 3D architecture.

7.3 Exploration Configuration

7.3.1 Parameters Exploration

We explore placement and routing options in the SoC Encounter in this design space exploration as shown in Table 25 and Table 26. We focus on timing and power optimization options in the 2D EDA tool to study how this 2D optimization process affects the 3D MPSoC architecture performance in terms of timing slack and power consumption. In addition, the chosen small number of options for the exploration is also because we have limited time to explore all other options since each exploration iteration for each tier requires about 4-5 hours of run time.

Table 25: EDA tool options for design space exploration

Placement options		Routing options	
Parameters	Values	Parameters	Values
Timing driven	False / true	Timing driven	False / true
Power driven	False / true	Route timing driven effort	5 (medium effort) / 10 (most aggressive)

7.3.2 Exploration Design Flow

Figure 88 shows the design flow used in this work to explore placement and routing options in the place and route tool. Synopsys Design Compiler was used for the logic synthesis while Cadence SoC Encounter was used for place and route of both tiers that is run in parallel during the exploration. 3D timing analysis and power analysis has been performed on the routed netlists of both tiers using Synopsys PrimeTime and PrimePower tool. The design space exploration is conducted using a combination of Shell and TCL scripts in Linux environment that automatically modifies the EDA tool options at each exploration iteration.

Table 26: Summary of design space exploration

Design ID	Placement options		Routing options	
	Timing Driven	Power Driven	Timing Driven	Route Timing Driven Effort
1	False	False	False	5
2	False	False	False	10
3	False	False	True	5
4	False	False	True	10
5	False	True	False	5
6	False	True	False	10
7	False	True	True	5
8	False	True	True	10
9	True	False	False	5
10	True	False	False	10
11	True	False	True	5
12	True	False	True	10
13	True	True	False	5
14	True	True	False	10
15	True	True	True	5
16	True	True	True	10

7.4 3D MPSoC Architectures for the Exploration

The 3D Mesh MPSoC architecture in this study is based on 16 processors with 3D NoC architecture as depicted in Figure 89 and Figure 90 for routed layout and tile floorplan respectively. The processor is based on the Openfire processor and the NoC architecture is based on 4x2 mesh topology with 3D router where both the processor and the NoC architectures have been explained in details in previous chapter. In order to provide more results on the impact of EDA tool on 3D architecture, we also conduct the exploration on heterogeneous 3D MPSoC architecture that has been described previosly in Chapter 6. In contrast with the 3D Mesh MPSoC architecture, this architecture has only 2D critical paths for both the processor as well as the NoC and therefore able to demonstrate the benefit of implementing 2D critical paths when designing 3D MPSoC architecture to take advantage of 2D optimization capability of the tool. For this exploration, we only focus on the timing performance and power consumption.

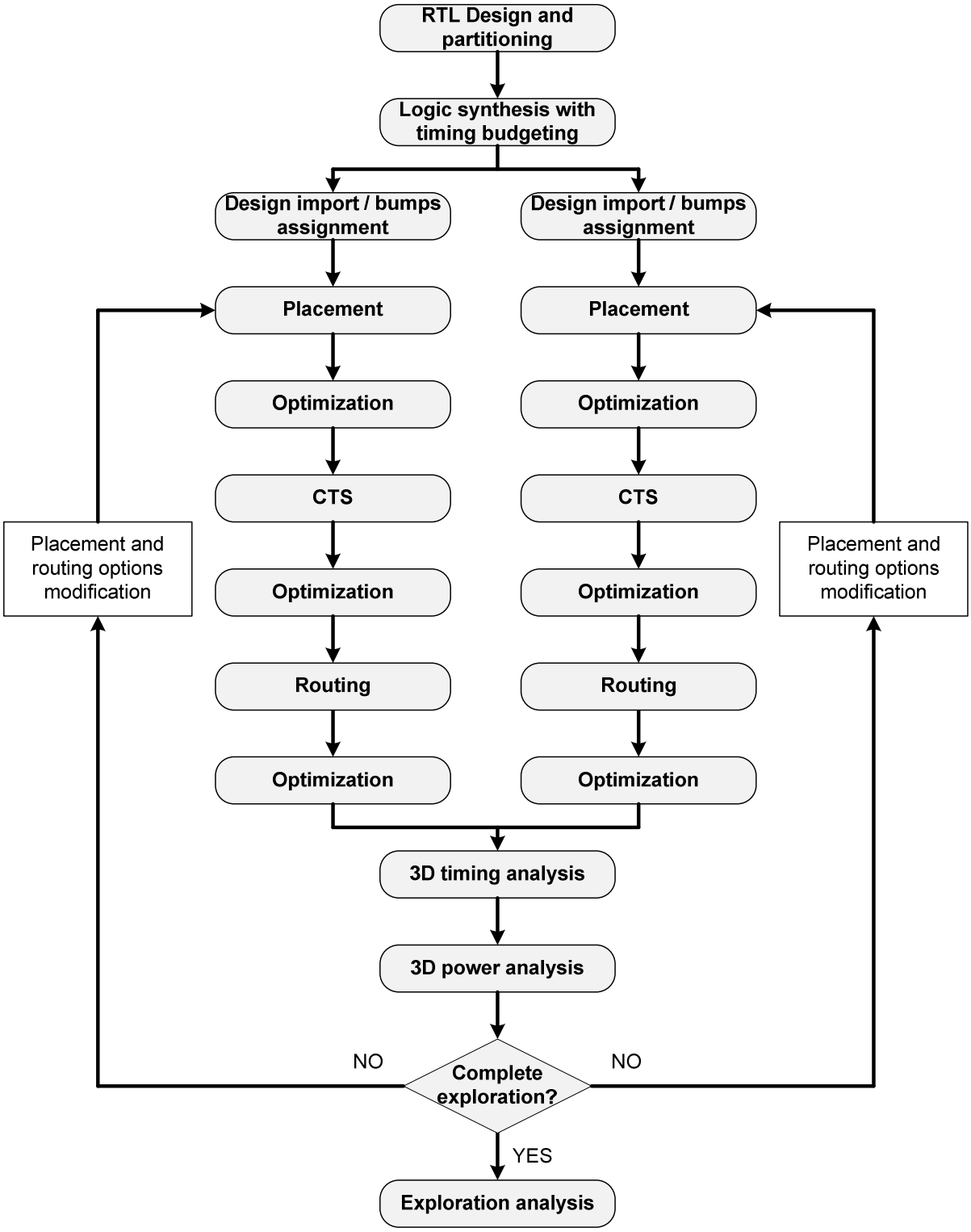


Figure 88: Design flow for EDA tool exploration

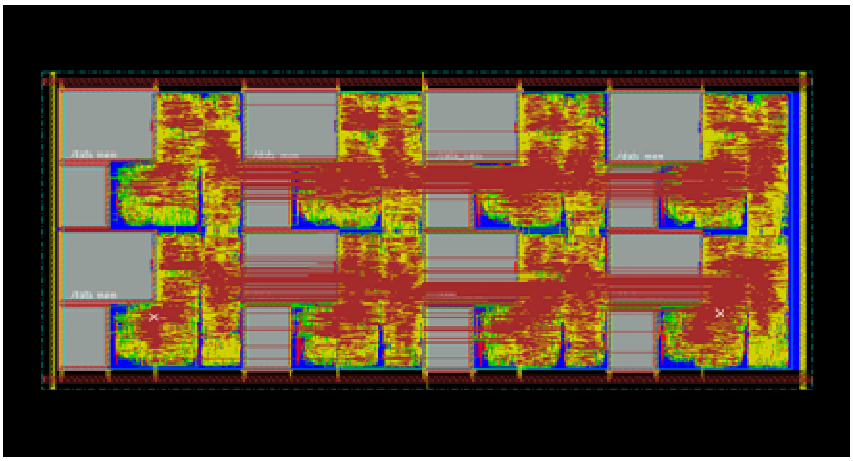


Figure 89: Bottom tier routed layout (top tier has the same layout)

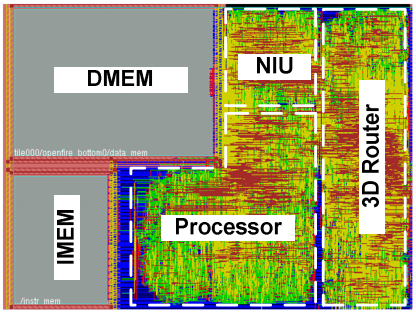


Figure 90: Close-up diagram of tile routed layout

Table 27: 3D architectures design summary for the exploration

Parameters	3D Mesh MPSoC	Heterogeneous 3D MPSoC
Core area (both tiers) (mm^2)	10.58	10.40
Total microbumps	595	3011
Microbumps per tile	74	188
Target clock period for NoC	3 ns	
Target clock period for processor	10 ns	

The design summary of for the 3D Mesh NoC and heterogeneous 3D stacking is shown in Table 27. Both architectures have almost similar core area but for total microbumps, heterogeneous 3D stacking has about five times more microbumps. The high number of microbumps for heterogeneous 3D stacking is due to the interface between NIU to the data memory and between processor to the instruction memory while for the 3D Mesh, the microbumps is only used by the routers' vertical ports. We use the same target clock period for the NoC and the processor for both architecture in order to have a fair comparison.

7.5 Exploration Results

In this section we discuss the exploration results based on the performance metrics which are the processor timing slack, NoC timing slack and power consumption.

7.5.1 Pocessor Timing Slack Analysis

For processor clock, the results from the exploration are shown in Figure 91 and Figure 92 for 3D Mesh MPSoC and heterogeneous 3D MPSoC respectively. The difference between the highest slack and lowest timing slack is about 2.9% for the 3D Mesh MPSoC while the value is reduced to 1.6% for the heterogeneous 3D MPSoC. Looking at the value of timing slack distribution for both graphs (y-axis), we clearly see that the timing slack is much lower for heterogeneous 3D MPSoC (maxiumum slack 0.16 ns) than for 3D Mesh NoC (maximum slack 0.4 ns). The reason is because for heterogeneous 3D MPSoC, the tile structure has been simplified (comparing the layouts of both 3D MPSoC architectures) due to the partitioning approach which separates the NoC architecture to the other tier (top tier). In contrast, the 3D Mesh MPSoC has a much density tile structure containing 3D router, NIU and processor components making it more difficult for the place and route tool (NanoRoute in SoC Encounter) to route the design due to higher complexity. In general, it can be concluded that 2D EDA tool options have a positive impact on the 2D timing performance of the 3D MPSoC architecture. In addition, it is shown that heterogeneous 3D MPSoC architecture has better timing performance than 3D Mesh MPSoC.

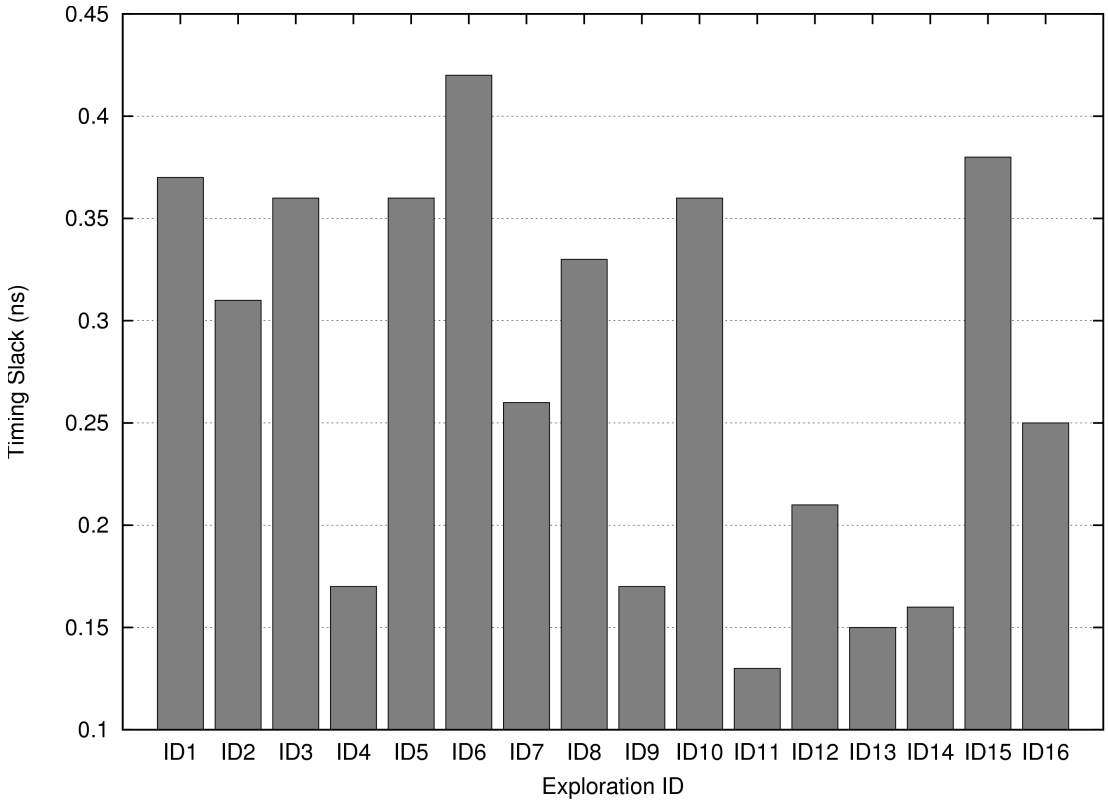


Figure 91: Processor timing slack (WNS) distribution for 3D Mesh MPSoC

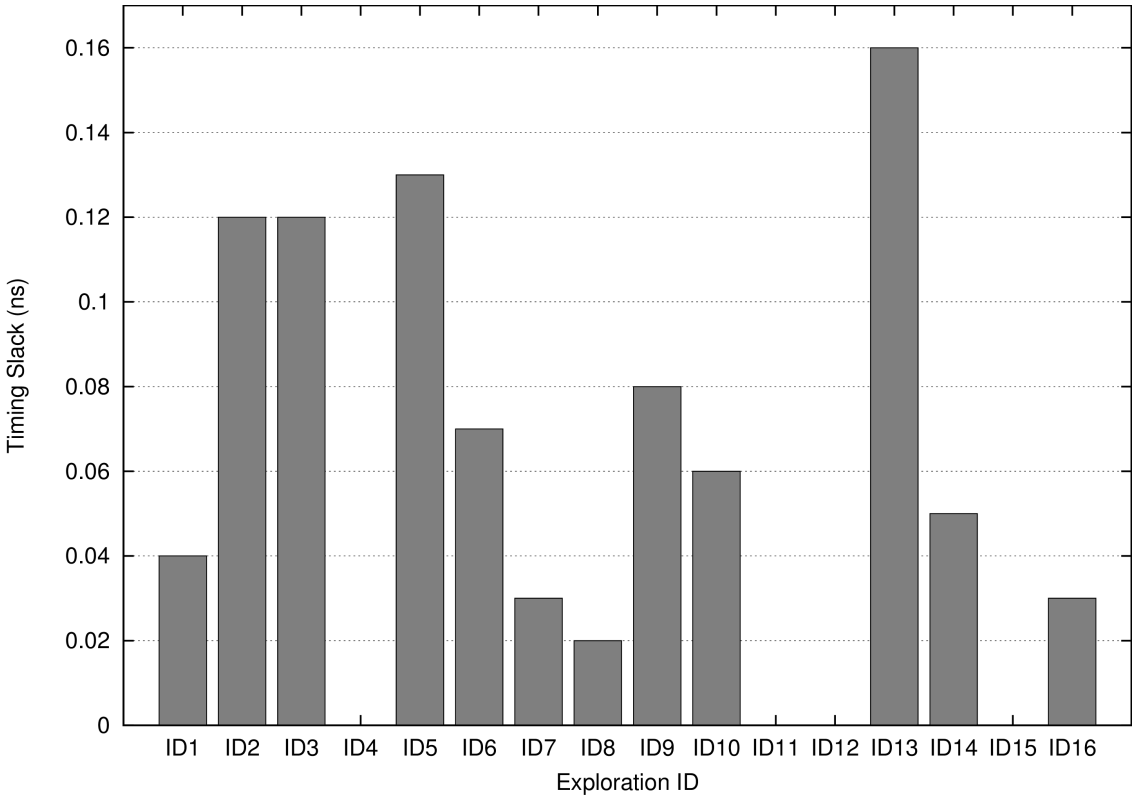


Figure 92: Processor Timing slack (WNS) distribution for heterogeneous 3D MPSoC

7.5.2 NoC Timing Slack Analysis

The results for NoC timing slack are shown in Figure 93 and Figure 94 for 3D Mesh MPSoC and heterogeneous 3D MPSoC respectively. For 3D Mesh MPSoC, the different between the highest and the lowest slack is about 13% but it is lower for the case of heterogeneous 3D MPSoC (about 7%), a reduction of 6%. For the 3D Mesh MPSoC, Exploration ID 15 shows the worst slack even though the timing-drven placement and timing-driven routing options have been used. This result suggests that the placement and routing options do not affect the 3D timing performance (3D Mesh MPSoC has 3D critical paths for NoC). Looking at the timing slack distrubtion values (y-axis) of both graphs, it is clearly shown that heterogeneous 3D MPSoC architecture has lower slack distribution (maximum slack 0.3 ns) than 3D Mesh MPSoC (maximum slack 1.75 ns). The reason for this high reduction is because heterogeneous 3D MPSoC architecture has 2D critical paths and thus the tool able to optimize it better by considering it as a normal 2D design. Moreover, the simplified tile structure on the top tier (NoC architecture) also contributes to this timing performance improvement which has been explained in the case of processor timing slack. In general, it can be concluded that 2D EDA tool options have a negative impact on the 3D timing performance of the 3D MPSoC architecture. Additionally, it has been shown that heterogeneous 3D MPSoC architecture has better timing performance than the 3D Mesh MPSoC architecture.

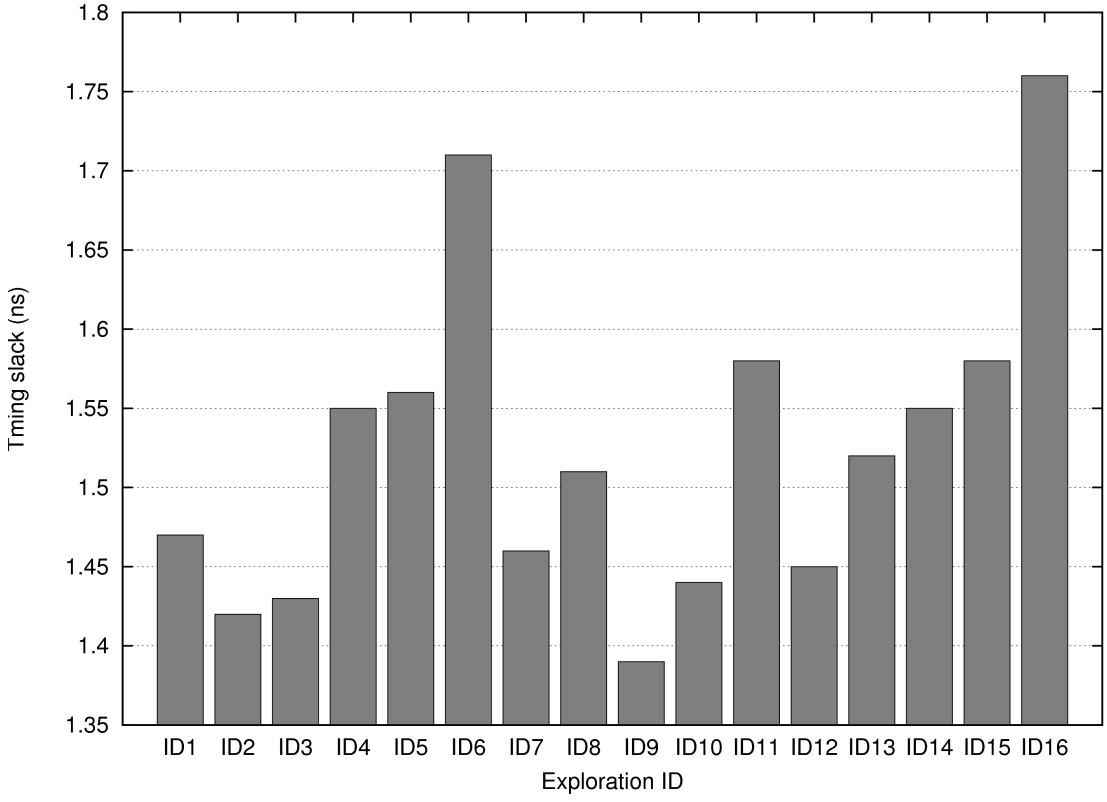


Figure 93: NoC timing slack (WNS) for 3D Mesh MPSoC

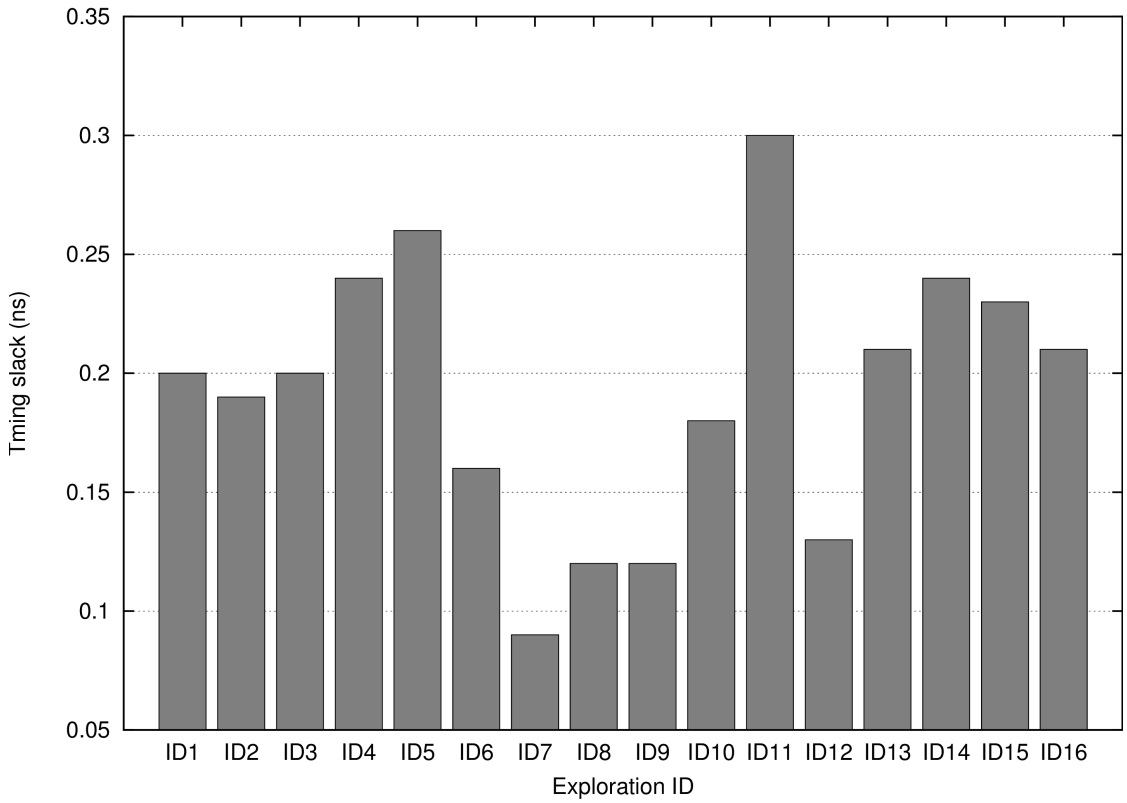


Figure 94: NoC timing slack (WNS) for heterogeneous 3D MPSoC

7.5.3 3D Power Consumption Analysis

The results for 3D power consumption are shown in Figure 95 and Figure 96 for 3D Mesh MPSoC and heterogeneous 3D MPSoC respectively. From these figures, it is clear shown that the 3D power consumption for both 3D MPSoC architectures does not varied very much which is about 40 mW between the highest and the lowest value in each graph. Using power driven in placement option reduces the total 3D power consumption as shown in ID5-ID8 and ID14-ID15 while using timing driven and power driven placement option produces the worst power consumption compared with other options for the 3D Mesh MPSoC. Considering the average power consumption value between both graphs, heterogeneous 3D MPSoC architecture has lower power than the 3D Mesh MPSoC (about 60 mW or 3% lower). In general, it can be concluded that 2D EDA tool options have no big impact on the power characteristic of 3D MPSoC architectures.

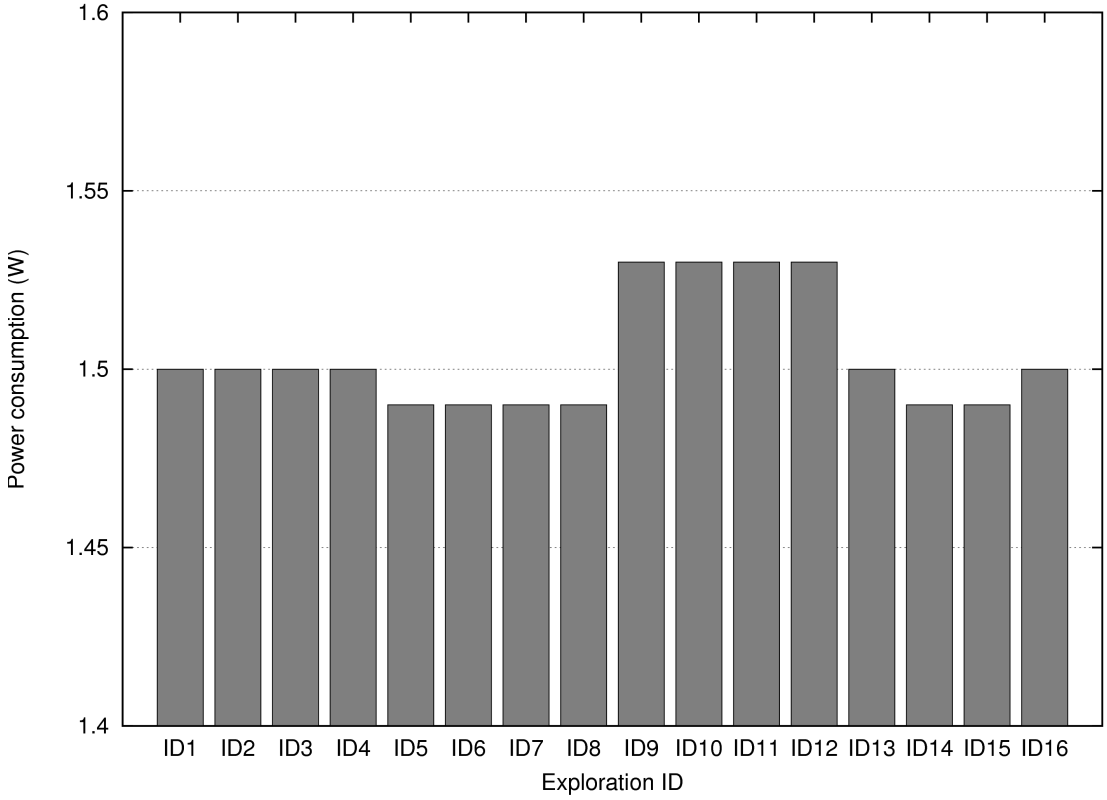


Figure 95: 3D power consumption for 3D Mesh MPSoC

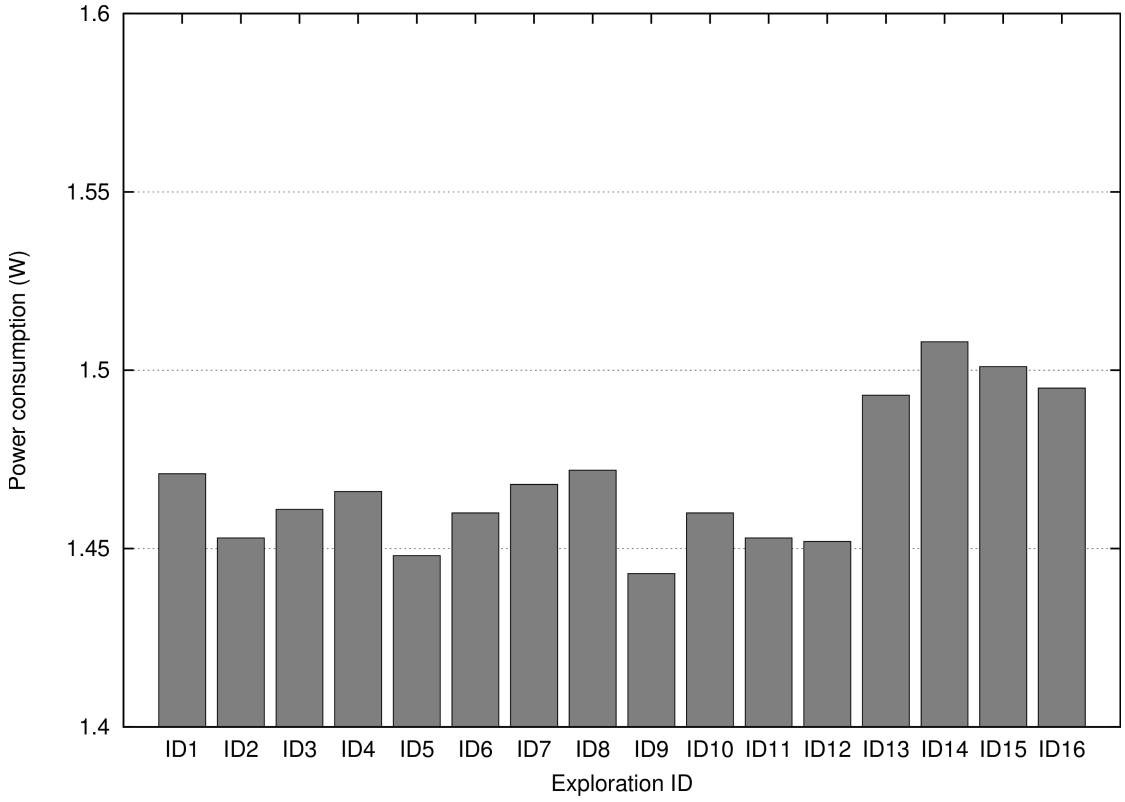


Figure 96: 3D power consumption for heterogeneous 3D MPSoC

7.6 Impact of using 2D EDA Tool on the Design of 3D MPSoCArchitecture

One of the important things that this design space exploration pointed out is that 2D EDA tool can be used optimizely when the 3D MPSoC architecture has 2D timing paths and the optimization process does not affect the 3D timing paths. This prove our finding in chapter 6 of heterogeneous 3D MPSoC stacking architecture which stating that we can optimize the 3D architecture by separating the design into several clock domains and put them in separate layers so that 2D tool can optimize it effectively as in normal 2D architecture. This approach is only feasible as long as there is no commercial 3D design tool available in the market with the capability of performing 3D design and optimization simultaneously.

Conclusion

In this chapter, we have presented a design space exploration of 2D EDA tool impact on the 3D MPSoC architecturs by analyzing the effect of different placement and routing options to the final 3D architecture timing and power characteristics. Results show that timing slack for both processor and NoC varied greatly than power consumption and total wirelength due to exploration option of timing driven properties in the place and route tool. Furthermore, it is shown that, as already pointed out in Chapter 6, to take benefits from 3D architecture as well as to fully utilize the capability of the state of the art 2D EDA tool to design 3D architecture, ensuring 2D critical paths rather than 3D critical paths in the target 3D architecture is one of the possible design approaches to be employed until the real 3D-aware design tool is commercially available.

CHAPTER 8
CONCLUSION AND FUTURE WORKS

8.1 Summary of Works

This work presented the study of 3D technology for multiprocessor with NoC architecture using physical design implementation based on Tezzaron 3D IC technology. Due to the lack of work in physical design implementation of 3D architectures to have more accurate performance analysis at hardware level, we carried out several design implementation experiments to better understand issues and design trade-offs in 3D technology from the architectural aspect as we have fixed the choice of technological aspect using Tezzaron 3D IC technology. Before starting to work on 3D IC technology, we have performed some experiments regarding NoC-based MPSoC implementation on FPGA to better understand the design issues related to the multiprocessor architecture. Running applications on the implemented MPSoC architecture allow us to analyze its performance especially parallel implementation performance which will be very useful as we are targeting to do performance analysis on the 3D MPSoC architecture on real ASIC implementation.

3D design flow is one of the important things to be considered due to unavailability of true 3D EDA tools to perform 3D synthesis, 3D place and route including 3D optimization for timing, power, thermal as well as other performance metrics. We have explained our 3D design flow that has been used in all the experiments in this thesis. The design flow is specific to two-tier Tezzaron 3D IC technology using face-to-face connection based on mirobumps structure but it is also applicable to any other 3D designs targeted to this technology. The main goal is to have early 3D performance estimation in particular timing verification as accurate as possible without having to complete the place and route step in order to save time from the timely iteration particularly for a very large design such 3D manycore architecture. Compared with prior design flows that have been proposed by several researchers, our design flow concentrate more on the 3D critical path analysis which eventually determine the performance of the 3D architecture in our case study.

We have explored 3D NoC architectures through physical design implementation. We found that 2D NoC topology (referring to the 3D Stacked Hexagonal NoC) have better performance than 3D NoC topology (3D Mesh NoC) following the results from physical implementations. Due to unequal inter-router wire length for 3D Stacked Mesh NoC, we proposed a new topology called hexagonal topology to improve performance by distributing the inter-router wire links equally. We have also conducted the effect of wire length on the 3D NoC architecture performance through the

implementation using two standard libraries from Global Foundries (130 nm) and ST Microelectronic (45 nm) representing old and advanced technologies. As wire length does not have significant effect to the performance (delay, power consumption) in 130 nm technology, the performance of 3D Stacked Hexagonal NoC is slightly worse than 3D Stacked Mesh NoC. However, in 45 nm tehnology, we can see that the speed of 3D Stacked Hexagonal NoC is slightly better than 3D Stacked Mesh NoC. This is because wire effect is becoming important in advanced technology and therefore equally distributed inter-router wire length in 3D Stacked Hexagonal NoC show a better performance. Because this design has relatively small area in 45 nm technology, the performance improvement of 3D Stacked Hexagonal NoC is not very pronounced when compared with other previous reported implementation that achieve quite significant performance improvement [153].

We have also implemented heterogeneous 3D stacking for NoC-based MPSoC architecture based on GALS approach in order to have further architectural exploration feasible to be built using 3D technology. GALS architecture provides better control for thermal and power management techniques and can reduce the effect of global clock tree structure. This heterogeneous 3D architecture uses separate clock domain for NoC and processor where the interface is handled by a dual clock FIFO structure built in inside the NIU block inside the NoC architecture. Through this experiment, we perform comparison analysis on the clock tree structure and critical paths between the 2D MPSoC and heterogeneous 3D MPSoC highlighting architectural trade-offs as well as issues related to 3D architecture design and implementation using 2D EDA tools. With respect to the 3D archictecture design using microbumps-based face-to-face stacking, even though microbumps does not suffer from the routing blockage and large area requirement as well as keep-out zone, we present issues highlighting restrictions on the microbumps pitch for designing more complex memory dominated 3D architecture when using 2D EDA tools thereby stressing to the need for real 3D-aware physical design tools to gain maximum benefits as well as to be able to perform 3D architecture explorations in meeting specific design requirements.

To study the impact of using 2D EDA tool on the performance of 3D architecture, we have performed a design space exploration of placement and routing options and analyze the 3D architecture performance in terms of timing slack (WNS), power consumption. As the main exploration options in the 2D EDA tool are timing driven option, therefore we observed that timing slack for both processor and NoC is varied more than power consumption results. The result of timing performance in this exploration proves the point of experiment in chapter 6 (heterogeneous 3D MPSoC stacking) that the use of 2D EDA tool does able to directly improve the timing

performance of designs with 3D critical paths. However, for 3D designs with 2D critical paths, the tool able to optimize the timing effectivey as in the normal 2D design flow suggesting that this methodology could be best suit for heterogeneous 3D architecture with GALS style implementation as functional blocks are separated in several clock domains. Although power driven in placement option has been chosen in the exploration, the impact to the 3D architecture performance is very little.

Finally, we have presented two MPSoC architectures developed by two teams, GIPSA-Lab in Grenoble (our team) and ENSTA ParisTech in Paris comparing the designs in terms of physical design implementations. While the number of vertical interconnections is not that much different between the two designs, both MPSoC architectures are differed mainly in the NoC topology used for the communication between processors which leads to the different 3D floorplan. Targeting for 3D fabrication through MPW services at CMP, the implementations allow us to analyze the performance of MPSoC architecture specifically the NoC in 3D architecture to be able to gain pratical understanding on the benefits of 3D technology. Moreover, this test chips also help us to identify design issues as well as trade-offs for 3D NoC architecture implementations.

8.2 Future Works

Following the experiments that have been conducted in this thesis, a number of subjects can be futher investigated. Some of the proposed future works are:

1. Explore various multi-stage network and mesh-based NoC topologies in 3D context to propose new topologies to dimension them according to wire delays and TSV delays. Mixed microbumps and TSV architecture can be considered for the physical design implementation for the performance analysis. Furthermore, a case study of more than two layer stacking can also be carried out to investigate the affect of the proposed topologies to the multilayer stacking.
2. Propose an algorithm for microbumps assignments for the inter-tiers signal connections considering both tier simultaneously in order to optimize vertical wirelengths as has been proposed by several works for TSV/microbumps assignments [163] [80] rather than using manual assignments which is not optimized. This could have pronounced effect to the overall 3D architecture performance because the 3D critical path can be fully optimized both vertically and horizontally.
3. Extend the experiments to evaluate the performance impact of multiple vertical connections

per signal compared with single vertical connection. Past studies have showed that using multiple TSVs (or vertical connections) have pronounced critical path delay reduction especially for large circuits co-design with the TSV placement optimization and design partitioning technique [153].

4. Perform the experiments for more than two tiers architecture to study the impact of multilayer stacks which can be done by taking two face-to-face wafers and stack them back-to-back for the Tezzaron 3D IC technology. It is also interested to perform the experiments using 3D technology using TSV in order to evaluate various issues and trade-off concering physical design implementation and performance impact to be compared with the microbump-based 3D technology.
5. Conduct a complete physical design analysis covering thermal implication, IR drop measurement, stress analysis and coupling noise analysis for signal as well as power/ground network to have complete understanding on the design and implementation of 3D multiprocessor architecture as have been considered in [124].

REFERENCES

[1] ITRS, “ITRS Report,” 2001. [Online]. Available: http://www.itrs.net.

[2] A. Roy, J. Xu, and M. H. Chowdhury, “Multi-core processors: A new way forward and challenges,” in *Microelectronics, 2008. ICM 2008. International Conference on*, 2008, pp. 454–457.

[3] L. Benini and G. De Micheli, *Networks on Chips: Technology And Tools*. Elsevier Morgan Kaufmann Publishers, 2006.

[4] D. Wentzlaff, P. Griffin, H. Hoffmann, L. Bao, B. Edwards, C. Ramey, M. Mattina, C.-C. Miao, J. F. Brown III, and A. Agarwal, “On-Chip Interconnection Architecture of the Tile Processor,” *Micro, IEEE*, vol. 27, no. 5, pp. 15–31, 2007.

[5] S. Borkar, “Design perspectives on 22nm CMOS and beyond,” in *Design Automation Conference, 2009. DAC ’09. 46th ACM/IEEE*, 2009, pp. 93–94.

[6] T. Ohba, N. Maeda, H. Kitada, K. Fujimoto, K. Suzuki, T. Nakamura, A. Kawai, and K. Arai, “Thinned wafer multi-stack 3DI technology,” *Microelectronic Engineering*, vol. 87, no. 3, pp. 485–490, 2010.

[7] S. Kosonocky, T. Burd, K. Kasprak, R. Schultz, and R. Stephany, “Designing in scaled technologies: 32 nm and beyond,” in *VLSI Technology (VLSIT), 2012 Symposium on*, 2012, pp. 147–148.

[8] J. U. Knickerbocker, P. S. Andry, B. Dang, R. R. Horton, M. J. Interrante, C. S. Patel, R. J. Polastre, K. Sakuma, R. Sirdeshmukh, E. J. Sprogis, S. M. Sri-Jayantha, A. M. Stephens, A. W. Topol, C. K. Tsang, B. C. Webb, and S. L. Wright, “Three-dimensional silicon integration,” *IBM Journal of Research and Development*, vol. 52, no. 6, pp. 553–569, 2008.

[9] W. Wolf, “Multiprocessor system-on-chip technology,” *Signal Processing Magazine, IEEE*, vol. 26, no. 6, pp. 50–54, 2009.

[10] J. Henkel, W. Wolf, and S. Chakradhar, “On-chip networks: a scalable, communication-centric embedded system design paradigm,” in *VLSI Design, 2004. Proceedings. 17th International Conference on*, 2004, pp. 845–851.

[11] “AMBA Open Specifications.” [Online]. Available: http://www.arm.com.

[12] “Wishbone Specification.” [Online]. Available: opencores.org.

[13] “STBus Communication System Concepts and Definitions.” [Online]. Available: www.st.com.

[14] “CoreConnect Bus Architecture.” [Online]. Available: www-01.ibm.com.

[15] L. Benini and G. De Micheli, “Networks on chips: a new SoC paradigm,” *Computer*, vol. 35, no. 1, pp. 70–78, 2002.

[16] C. A. Zeferino, M. E. Kreutz, L. Carro, and A. A. Susin, "A study on communication issues for systems-on-chip," in *Integrated Circuits and Systems Design, 2002. Proceedings. 15th Symposium on*, 2002, pp. 121–126.

[17] A. K. Lusala, P. Manet, B. Rousseau, and J.-D. Legat, "NoC Implementation in FPGA using Torus Topology," in *Field Programmable Logic and Applications, 2007. FPL 2007. International Conference on*, 2007, pp. 778–781.

[18] G. Du, D. Zhang, Y. Song, M. Gao, L. Geng, and N. Hou, "Scalability Study on Mesh Based Network on Chip," in *Computational Intelligence and Industrial Application, 2008. PACIIA '08. Pacific-Asia Workshop on*, 2008, vol. 2, pp. 681–685.

[19] G. Luo-Feng, D. Gao-ming, Z. Duo-Li, G. Ming-Lun, H. Ning, and S. Yu-Kun, "Design and performance evaluation of a 2D-mesh Network on Chip prototype using FPGA," in *Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on*, 2008, pp. 1264–1267.

[20] T. Le and M. Khalid, "NoC prototyping on FPGAs: A case study using an image processing benchmark," in *Electro/Information Technology, 2009. eit '09. IEEE International Conference on*, 2009, pp. 441–445.

[21] C. Seiculescu, S. Murali, L. Benini, and G. De Micheli, "SunFloor 3D: A Tool for Networks on Chip Topology Synthesis for 3-D Systems on Chips," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 29, no. 12, pp. 1987–2000, 2010.

[22] A. Adriahantenaina, H. Charlery, A. Greiner, L. Mortiez, and C. A. Zeferino, "SPIN: a scalable, packet switched, on-chip micro-network," in *Design, Automation and Test in Europe Conference and Exhibition, 2003*, 2003, pp. 70–73 suppl.

[23] H. Nikolov, M. Thompson, T. Stefanov, A. Pimentel, S. Polstra, R. Bose, C. Zissulescu, and E. Deprettere, "Daedalus: Toward composable multimedia MP-SoC design," in *Design Automation Conference, 2008. DAC 2008. 45th ACM/IEEE*, 2008, pp. 574–579.

[24] J. Wu, J. Williams, N. Bergmann, and P. Sutton, "Design Exploration for FPGA-Based Multiprocessor Architecture: JPEG Encoding Case Study," in *Field Programmable Custom Computing Machines, 2009. FCCM '09. 17th IEEE Symposium on*, 2009, pp. 299–302.

[25] J. S. Patrick, J. L. Sanders, L. S. DeBrunner, V. E. DeBrunner, and S. Radharkrishnan, "JPEG compression/decompression via parallel processing," in *Signals, Systems and Computers, 1996. Conference Record of the Thirtieth Asilomar Conference on*, 1996, vol. 1, pp. 596–600 vol.1.

[26] "Arteris NoCCompiler." [Online]. Available: www.arteris.com.

[27] "Platform Studio and Xilinx Embedded Developement Kit (EDK)," 2012. [Online]. Available: www.xilinx.com.

[28] "EVE." [Online]. Available: www.eve-team.com.

[29] X. Li and O. Hammami, "An Automatic Design Flow for Data Parallel and Pipelined Signal Processing Applications on Embedded Multiprocessor with NoC: Application to

Cryptography," *International Journal of Reconfigurable Computing*, vol. 2009, pp. 1–14, 2009.

[30] "OCP-IP." [Online]. Available: www.ocp-ip.com.

[31] Xilinx, "MicroBlaze Soft Processor Core," 2012. [Online]. Available: http://www.xilinx.com/tools/microblaze.htm.

[32] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete Cosine Transform," *Computers, IEEE Transactions on*, vol. C–23, no. 1, pp. 90–93, 1974.

[33] Q. Liu, R. J. Sclabassi, and M. Sun, "A DCT-Domain Approach to Image Change Detection and Its Application to Patient Video Monitoring," in *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, 2005, pp. 1–4.

[34] D. Chikouche, R. Benzid, and M. Bentoumi, "Application of the DCT and Arithmetic Coding to Medical Image Compression," in *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, 2008, pp. 1–5.

[35] T. N. Theis, "Future of interconnection technology," *IBM Journal of Research and Development*, vol. 44, no. 3, pp. 379–390, 2000.

[36] J. Howard, S. Dighe, S. R. Vangal, G. Ruhl, N. Borkar, S. Jain, V. Erraguntla, M. Konow, M. Riepen, M. Gries, G. Droege, T. Lund-Larsen, S. Steibl, S. Borkar, V. K. De, and R. Van Der Wijngaart, "A 48-Core IA-32 Processor in 45 nm CMOS Using On-Die Message-Passing and DVFS for Performance and Power Scaling," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 1, pp. 173–183, 2011.

[37] P. Kapur, "Scaling Induced Performance Challenges/Limitations Of On-Chip Metal Interconnects And Comparisons With Optical Interconnects," PhD Thesis, Stanford University, 2002.

[38] "ITRS Report Interconnect," 2011. [Online]. Available: http://www.itrs.net/home.html.

[39] T. N. Theis, "Challenges in the extension of hierarchical wiring systems," *Proceedings - Electrochemical Society*, vol. 6, pp. 1–11, 1999.

[40] S. S. Iyer, G. Freeman, C. Brodsky, A. I. Chou, D. Corliss, S. H. Jain, N. Lustig, V. McGahay, S. Narasimha, J. Norum, K. A. Nummy, P. Parries, S. Sankaran, C. D. Sheraw, P. R. Varanasi, G. Wang, M. E. Weybright, X. Yu, E. Crabbe, and P. Agnello, "45-nm silicon-on-insulator CMOS technology integrating embedded DRAM for high-performance server and ASIC applications," *IBM Journal of Research and Development*, vol. 55, no. 3, pp. 5:1–5:14, 2011.

[41] K.-H. Koo, P. Kapur, and K. C. Saraswat, "Compact Performance Models and Comparisons for Gigascale On-Chip Global Interconnect Technologies," *Electron Devices, IEEE Transactions on*, vol. 56, no. 9, pp. 1787–1798, 2009.

[42] Y. Akasaka and T. Nishimura, "Concept and basic technologies for 3-D IC structure," in *Electron Devices Meeting, 1986 International*, 1986, vol. 32, pp. 488–491.

[43] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon, "Demystifying 3D ICs: the pros and cons of going vertical," *Design & Test of Computers, IEEE*, vol. 22, no. 6, pp. 498–510, 2005.

[44] E. Beyne, "The rise of the 3rd dimension for system intergration," in *Interconnect Technology Conference, 2006 International*, 2006, pp. 1–5.

[45] A. Fazzi, L. Magagni, M. Mirandola, B. Charlet, L. Di Cioccio, E. Jung, R. Canegallo, and R. Guerrieri, "3-D Capacitive Interconnections for Wafer-Level and Die-Level Assembly," *Solid-State Circuits, IEEE Journal of*, vol. 42, no. 10, pp. 2270–2282, 2007.

[46] M. Saen, K. Osada, Y. Okuma, K. Niitsu, Y. Shimazaki, Y. Sugimori, Y. Kohama, K. Kasuga, I. Nonomura, N. Irie, T. Hattori, A. Hasegawa, and T. Kuroda, "3-D System Integration of Processor and Multi-Stacked SRAMs Using Inductive-Coupling Link," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 4, pp. 856–862, 2010.

[47] H.-C. Lu, G.-M. Wu, C. Pan, and Y.-T. Chou, "Coupling coefficient improvement for inductor coupled vertical interconnect in 3D IC die stacking," in *Electronic Components and Technology Conference, 2009. ECTC 2009. 59th*, 2009, pp. 1207–1212.

[48] M. Koyanagi, T. Fukushima, and T. Tanaka, "High-Density Through Silicon Vias for 3-D LSIs," *Proceedings of the IEEE*, vol. 97, no. 1, pp. 49–59, 2009.

[49] V. F. Pavlidis and E. Friedman, *Three-Dimensional Integrated Circuit Design*. Morgan Kaufmann, 2009.

[50] H. Kondo, S. Otani, M. Nakajima, O. Yamamoto, N. Masui, N. Okumura, M. Sakugawa, M. Kitao, K. Ishimi, M. Sato, F. Fukuzawa, S. Imasu, N. Kinoshita, Y. Ota, K. Arimoto, and T. Shimizu, "Heterogeneous Multicore SoC With SiP for Secure Multimedia Applications," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 8, pp. 2251–2259, 2009.

[51] K. C. Saraswat, "3-D ICs: Motivation, performance analysis, technology and applications," in *Physical and Failure Analysis of Integrated Circuits (IPFA), 2010 17th IEEE International Symposium on the*, 2010, pp. 1–6.

[52] S. Wong, A. El-Gamal, P. Griffin, Y. Nishi, F. Pease, and J. Plummer, "Monolithic 3D Integrated Circuits," in *VLSI Technology, Systems and Applications, 2007. VLSI-TSA 2007. International Symposium on*, 2007, pp. 1–4.

[53] O. Thomas, M. Vinet, O. Rozeau, P. Batude, and A. Valentian, "Compact 6T SRAM cell with robust read/write stabilizing design in 45nm Monolithic 3D IC technology," in *IC Design and Technology, 2009. ICICDT '09. IEEE International Conference on*, 2009, pp. 195–198.

[54] C. Liu and S. K. Lim, "A design tradeoff study with monolithic 3D integration," in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, 2012, pp. 529–536.

[55] R. Chaware, K. Nagarajan, and S. Ramalingam, "Assembly and reliability challenges in 3D integration of 28nm FPGA die on a large high density 65nm passive interposer," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, 2012, pp. 279–283.

[56] N. Kim, D. Wu, J. Carrel, J.-H. Kim, and P. Wu, "Channel design methodology for 28Gb/s SerDes FPGA applications with stacked silicon interconnect technology," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, 2012, pp. 1786–1793.

[57] H.-E. Kim, J.-S. Yoon, K.-D. Hwang, Y.-J. Kim, J.-S. Park, and L.-S. Kim, "A 275mW heterogeneous multimedia processor for IC-stacking on Si-interposer," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 128–130.

[58] P. G. Emma and E. Kursun, "Opportunities and Challenges for 3D Systems and Their Design," *Design & Test of Computers, IEEE*, vol. 26, no. 5, pp. 6–14, 2009.

[59] T. Thorolfsson, P. D. Franzon, and K. Gonsalves, "Design automation for a 3DIC FFT processor for synthetic aperture radar: A case study," in *Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE*, 2009, pp. 51–56.

[60] B. Vaidyanathan, W.-L. Hung, F. Wang, Y. Xie, V. Narayanan, M. J. Irwin, and M. Gsrc, "Architecting Microprocessor Components in 3D Design Space," in *VLSI Design, 2007. Held jointly with 6th International Conference on Embedded Systems., 20th International Conference on*, 2007, pp. 103–108.

[61] Y. Xie, "Processor Architecture Design Using 3D Integration Technology," *2010 23rd International Conference on VLSI Design*, pp. 446–451, Jan. 2010.

[62] R. Weerasekera, M. Grange, D. Pamunuwa, H. Tenhunen, and L.-R. Zheng, "Compact modelling of Through-Silicon Vias (TSVs) in three-dimensional (3-D) integrated circuits," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, 2009, pp. 1–8.

[63] T. Hsu, K. Chiang, J.-Y. Lai, and Y.-P. Wang, "Electrical characterization of through silicon via (TSV) for high-speed memory application," in *Electronic Manufacturing Technology Symposium (IEMT), 2008 33rd IEEE/CPMT International*, 2008, pp. 1–5.

[64] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs," *Electron Devices, IEEE Transactions on*, vol. 57, no. 1, pp. 256–262, 2010.

[65] J. Ouyang, G. Sun, D. Y. Chen, L. Duan, T. Zhang, Y. Xie, and M. J. Irwin, "Arithmetic unit design using 180nm TSV-based 3D stacking technology," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, 2009, pp. 1–4.

[66] P. Jacob, A. Zia, O. Erdogan, P. M. Belemjian, J.-W. Kim, M. Chu, R. P. Kraft, J. F. Mcdonald, and K. Bernstein, "Mitigating Memory Wall Effects in High-Clock-Rate and Multicore CMOS 3-D Processor Memory Stacks," *Proceedings of the IEEE*, vol. 97, no. 1, pp. 108–122, 2009.

[67] J.-S. Kim, C. S. Oh, H. Lee, D. Lee, H. R. Hwang, S. Hwang, B. Na, J. Moon, J.-G. Kim, H. Park, J.-W. Ryu, K. Park, S. K. Kang, S.-Y. Kim, H. Kim, J.-M. Bang, H. Cho, M. Jang, C. Han, J.-B. LeeLee, J. S. Choi, and Y.-H. Jun, "A 1.2 V 12.8 GB/s 2 Gb Mobile Wide-I/O DRAM With 4 x 128 I/Os Using TSV Based Stacking," *Solid-State Circuits, IEEE Journal of*, vol. 27, no. 1, pp. 107–116, 2011.

[68] A. K. Coskun, A. B. Kahng, and T. S. Rosing, “Temperature- and Cost-Aware Design of 3D Multiprocessor Architectures,” in *2009 12th Euromicro Conference on Digital System Design, Architectures, Methods and Tools*, 2009, pp. 183–190.

[69] C.-L. Huang, N.-S. Chang, C.-S. Chen, C.-P. Lin, C.-M. Wu, and C.-M. Huang, “A novel design methodology for hybrid process 3D-IC,” in *VLSI Design, Automation, and Test (VLSI-DAT), 2012 International Symposium on*, 2012, pp. 1–4.

[70] V. K. Jain, S. Bhanja, G. H. Chapman, and L. Doddannagari, “A highly reconfigurable computing array: DSP plane of a 3D heterogeneous SoC,” in *SOC Conference, 2005. Proceedings. IEEE International*, 2005, pp. 243–246.

[71] S. K. Kim, C. C. Liu, L. Xue, and S. Tiwari, “Crosstalk reduction in mixed-signal 3-D integrated circuits with interdevice layer ground planes,” *Electron Devices, IEEE Transactions on*, vol. 52, no. 7, pp. 1459–1467, 2005.

[72] D. H. Triyoso, T. B. Dao, T. Kropewnicki, F. Martinez, R. Noble, and M. Hamilton, “Progress and challenges of tungsten-filled through-silicon via,” in *IC Design and Technology (ICICDT), 2010 IEEE International Conference on*, 2010, pp. 118–121.

[73] M. J. Wolf, T. Dretschkow, B. Wunderle, N. Jurgensen, G. Engelmann, O. Ehrmann, A. Uhlig, B. Michel, and H. Reichl, “High aspect ratio TSV copper filling with different seed layers,” in *Electronic Components and Technology Conference, 2008. ECTC 2008. 58th*, 2008, pp. 563–570.

[74] G. Katti, A. Mercha, J. Van Olmen, C. Huyghebaert, A. Jourdain, M. Stucchi, M. Rakowski, I. Debusschere, P. Soussan, W. Dehaene, K. De Meyer, Y. Travaly, E. Beyne, S. Biesemans, and B. Swinnen, “3D stacked ICs using Cu TSVs and Die to Wafer Hybrid Collective bonding,” in *Electron Devices Meeting (IEDM), 2009 IEEE International*, 2009, pp. 1–4.

[75] T. Jiang and S. Luo, “3D Integration-Present and Future,” in *Electronics Packaging Technology Conference, 2008. EPTC 2008. 10th*, 2008, pp. 373–378.

[76] K. Nomura, K. Abe, S. Fujita, Y. Kurosawa, and A. Kageshima, “Performance Analysis of 3D-IC for Multi-Core Processors in sub-65nm CMOS technologies,” in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2010, pp. 2876–2879.

[77] R. Yarema, “The Via Revolution,” in *19th International Workshop on Vertex Dectectors - VERTEX 2010*, 2010, pp. 1–11.

[78] R. S. Patti, “Three-Dimensional Integrated Circuits and the Future of System-on-Chip Designs,” *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1214–1224, 2006.

[79] M. Rimskog and T. Bauer, “High density Through Silicon Via (TSV),” in *Design, Test, Integration and Packaging of MEMS/MOEMS, 2008. MEMS/MOEMS 2008. Symposium on*, 2008, pp. 105–108.

[80] D. H. Kim, K. Athikulwongse, and S. K. Lim, “A study of through-silicon-via impact on the 3D stacked IC layout,” in *Proceedings of the 2009 International Conference on Computer-Aided Design*, 2009, pp. 674–680.

[81] E. C. Oh and P. D. Franzon, "Technology impact analysis for 3D TCAM," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, 2009, pp. 1–5.

[82] A. W. Topol, D. C. La Tulipe, L. Shi, D. J. Frank, K. Bernstein, S. E. Steen, A. Kumar, G. U. Singco, A. M. Young, K. W. Guarini, and M. Ieong, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 491–506, 2006.

[83] D. Fick, R. G. Dreslinski, B. Giridhar, G. Kim, S. Seo, M. Fojtik, S. Satpathy, Y. Lee, D. Kim, N. Liu, M. Wieckowski, G. Chen, T. Mudge, D. Sylvester, and D. Blaauw, "Centip3De: A 3930DMIPS/W configurable near-threshold 3D stacked system with 64 ARM Cortex-M3 cores," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, 2012, pp. 190–192.

[84] C. Mineo, R. Jenkal, S. Melamed, and W. R. Davis, "Inter-die signaling in three dimensional integrated circuits," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, 2008, pp. 655–658.

[85] P. Gueguen, C. Ventosa, L. Di Cioccio, H. Moriceau, F. Grossi, M. Rivoire, P. Leduc, and L. Clavelier, "Physics of direct bonding: Applications to 3D heterogeneous or monolithic integration," *Microelectronic Engineering*, vol. 87, no. 3, pp. 477–484, Mar. 2010.

[86] C.-T. Ko and K.-N. Chen, "Wafer-level bonding/stacking technology for 3D integration," *Microelectronics Reliability*, vol. 50, no. 4, pp. 481–488, Apr. 2010.

[87] L. Di Cioccio, I. Radu, P. Gueguen, and M. Sadaka, "Direct bonding for wafer level 3D integration," in *IC Design and Technology (ICICDT), 2010 IEEE International Conference on*, 2010, pp. 110–113.

[88] J. J. McMahon, E. Chan, S. H. Lee, R. J. Gutmann, and J.-Q. Lu, "Bonding interfaces in wafer-level metal/adhesive bonded 3D integration," in *Electronic Components and Technology Conference, 2008. ECTC 2008. 58th*, 2008, pp. 871–878.

[89] M. Nimura, J. Mizuno, K. Sakuma, and S. Shoji, "Solder/adhesive bonding using simple planarization technique for 3D integration," in *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st*, 2011, pp. 1147–1152.

[90] F. Liu, R. R. Yu, A. M. Young, J. P. Doyle, X. Wang, L. Shi, K.-N. Chen, X. Li, D. A. Dipaola, D. Brown, C. T. Ryan, J. A. Hagan, K. H. Wong, M. Lu, X. Gu, N. R. Klymko, E. D. Perfecto, A. G. Merryman, K. A. Kelly, S. Purushothaman, S. J. Koester, R. Wisnieff, and W. Haensch, "A 300-mm wafer-level three-dimensional integration scheme using tungsten through-silicon via and hybrid Cu-adhesive bonding," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 1–4.

[91] K. Sakuma, "Development of 3D Chip Integration Technology," in *Nano-Semiconductors: Devices and Technology*, K. Iniewski, Ed. CRC Press, 2011, pp. 174–221.

[92] S. Farrens, "Wafer and Die Bonding Technologies for 3D Integration," *Materials Research Society Symposium Proceedings*, vol. 1112, no. 1, pp. 55–65, 2008.

[93] G. H. Loh, "Three-Dimensional Microprocessor Design," in *Three Dimensional Integrated Circuit Design: EDA, Design and Microarchitectures*, Y. Xie, J. Cong, and S. Sapatnekar, Eds. Springer, 2010, pp. 161–188.

[94] “IMEC.” [Online]. Available: http://www2.imec.be/be_en/research.html.

[95] “Ziptronix.” [Online]. Available: http://www.ziptronix.com/.

[96] “MIT Lincoln Lab.” [Online]. Available: http://www.ll.mit.edu/index.html.

[97] “BeSang.” [Online]. Available: http://www.besang.com.

[98] S. Lee and D. K. Schroder, “3D IC architecture for high density memories,” in *Memory Workshop (IMW), 2010 IEEE International*, 2010, pp. 1–6.

[99] “Tezzaron 3D IC technology.” [Online]. Available: http://www.tezzaron.com/.

[100] K. Kuhn, C. Kenyon, A. Kornfeld, M. Liu, A. Maheshwari, W.-K. Shih, S. Sivakumar, G. Taylor, P. VanDerVoorn, and K. Zawadzki, “Managing Process Variation in Intel’s 45nm CMOS Technology,” *Intel Technology Journal*, vol. 12, no. 02, pp. 93–110, 2008.

[101] K. Kuhn, “CMOS scaling beyond 32nm: Challenges and opportunities,” in *Design Automation Conference, 2009. DAC ’09. 46th ACM/IEEE*, 2009, pp. 310–313.

[102] D. Velenis, M. Stucchi, E. J. Marinissen, B. Swinnen, and E. Beyne, “Impact of 3D design choices on manufacturing cost,” in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, 2009, pp. 1–5.

[103] B. Black, M. Annavaram, N. Brekelbaum, J. Devale, L. Jiang, G. H. Loh, D. McCauley, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb, “Die Stacking (3D) Microarchitecture,” in *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*, 2006, pp. 469–479.

[104] R. Weerasekera, L. Zheng, D. Pamunuwa, and H. Tenhunen, “Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs,” in *Computer-Aided Design, 2007. ICCAD 2007. IEEE/ACM International Conference on*, 2007, pp. 212–219.

[105] Y. J. Park, M. Zeng, B. Lee, J.-A. Lee, S. G. Kang, and C. H. Kim, “Thermal Analysis for 3D Multi-core Processors with Dynamic Frequency Scaling,” in *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*, 2010, pp. 69–74.

[106] D. Li, S. X. S. X.-D. Tan, E. H. Pacheco, and M. Tirumala, “Architecture-Level Thermal Characterization for Multicore Microprocessors,” *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 10, pp. 1495–1507, 2009.

[107] M. Santarini, “Thermal integrity: a must for low-power-IC digital design,” *EDN*, 2005. [Online]. Available: www.edn.com.

[108] K. Puttaswamy and G. H. Loh, “Thermal Herding: Microarchitecture Techniques for Controlling Hotspots in High-Performance 3D-Integrated Processors,” in *High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on*, 2007, pp. 193–204.

[109] E. Wong and S. K. Lim, “3D Floorplanning with Thermal Vias,” in *Design, Automation and Test in Europe, 2006. DATE’06. Proceedings*, 2006, vol. 1, pp. 1–6.

[110] W. Hung, G. M. Link, Y. Xie, N. Vijaykrishnan, and M. J. Irwin, "Interconnect and thermal-aware floorplanning for 3D microprocessors," in *Quality Electronic Design, 2006. ISQED '06. 7th International Symposium on*, 2006, pp. 1–6.

[111] J. Yang, K. Athikulwongse, Y. Lee, S. K. Lim, and D. Z. Pan, "TSV stress aware timing analysis with applications to 3D-IC layout optimization," in *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, 2010, pp. 803–806.

[112] M. C. Hsieh, Y.-Y. Hsu, and C.-L. Chang, "Thermal Stress Analysis of Cu/Low-k Interconnects in 3D-IC Structures," in *Microsystems, Packaging, Assembly Conference Taiwan, 2006. IMPACT 2006. International*, 2006, pp. 1–4.

[113] H. K. Lu, "Thermo-mechanical Reliability of 3-D Interconnects Containing Through-Silicon-Vias (TSVs)," 2010.

[114] A. P. Karmarkar, "Performance and reliability analysis of 3D-integration structures employing Through Silicon Via (TSV)," in *Reliability Physics Symposium (IRPS), 2009 IEEE International*, 2009, pp. 682–687.

[115] N. H. Khan, S. M. Alam, and S. Hassoun, "Power Delivery Design for 3-D ICs Using Different Through-Silicon Via (TSV) Technologies," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 4, pp. 647–658, 2011.

[116] M. B. Healy and S. K. Lim, "A novel TSV topology for many-tier 3D power-delivery networks," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011*, 2011, pp. 1–4.

[117] P. D. Franzon, W. R. Davis, and T. Thorolffson, "Creating 3D specific systems: Architecture, design and CAD," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010*, 2010, pp. 1684–1688.

[118] M. Gupta, G. Rajagopalan, C. K. Hong, J.-Q. Lu, K. Rose, and R. J. Gutmann, "Planarization yield limiters for wafer-scale 3D ICs," in *Advanced Semiconductor Manufacturing 2002 IEEE/SEMI Conference and Workshop*, 2002, pp. 278–283.

[119] D. V Campbell, "Yield modeling of 3D integrated wafer scale assemblies," in *Electronic Components and Technology Conference (ECTC), 2010 Proceedings 60th*, 2010, pp. 1935–1938.

[120] G. Smith, L. Smith, S. Hosali, and S. Arkalgud, "Yield considerations in the choice of 3D technology," in *Semiconductor Manufacturing, 2007. ISSM 2007. International Symposium on*, 2007, pp. 1–3.

[121] E. J. Marinissen, "Testing TSV-based three-dimensional stacked ICs," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010*, 2010, pp. 1689–1694.

[122] H.-H. S. Lee and K. Chakrabarty, "Test Challenges for 3D Integrated Circuits," *Design & Test of Computers, IEEE*, vol. 26, no. 5, pp. 26–35, 2009.

[123] Q. Gu, Z. Xu, J. Ko, and M.-C. F. Chang, "Two 10Gb/s/pin Low-Power Interconnect Methods for 3D ICs," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2007 IEEE International*, 2007, pp. 448–614.

[124] M. B. Healy, K. Athikulwongse, R. Goel, M. Hossain, D. H. Kim, Y.-J. Lee, D. L. Lewis, T.-W. Lin, C. Liu, M. Jung, B. Ouellette, M. Pathak, H. Sane, G. Shen, D. H. Woo, X. Zhao, G. H. Loh, H.-H. S. Lee, and S. K. Lim, "Design and analysis of 3D-MAPS: A many-core 3D processor with stacked memory," in *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, 2010, pp. 1–4.

[125] T. Zhang, K. Wang, Y. Feng, Y. Chen, Q. Li, B. Shao, J. Xie, X. Song, L. Duan, Y. Xie, X. Cheng, and Y.-L. Lin, "A 3D SoC design for H.264 application with on-chip DRAM stacking," in *3D Systems Integration Conference (3DIC), 2010 IEEE International*, 2010, pp. 1–6.

[126] I. Loi, P. Marchal, A. Pullini, and L. Benini, "3D NoCs - Unifying inter & intra chip communication," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 2010, pp. 3337–3340.

[127] X. Chen, T. Zhu, and W. R. Davis, "Three-dimensional SRAM design with on-chip access time measurement," *Electronics Letters*, vol. 47, no. 8, p. 485, 2011.

[128] L. Zhou, C. Wakayama, N. Jangkrajarng, B. Hu, and C.-J. R. Shi, "A high-throughput low-power fully parallel 1024-bit 1/2 -rate low density parity check code decoder in 3D integrated circuits," in *Design Automation, 2006. Asia and South Pacific Conference on*, 2006, pp. 92–93.

[129] J. Xie, X. Dong, and Y. Xie, "3D memory stacking for fast checkpointing/restore applications," in *3D Systems Integration Conference (3DIC), 2010 IEEE International*, 2010, pp. 1–6.

[130] G. Beanato, P. Giovannini, A. Cevrero, P. Athanasopoulos, M. Zervas, Y. Temiz, and Y. Leblebici, "Design and Testing Strategies for Modular 3-D-Multiprocessor Systems Using Die-Level Through Silicon Via Technology," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 2, no. 2, pp. 295–306, 2012.

[131] H. Hua, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, R. Jenkal, and W. R. Davis, "Exploring compromises among timing, power and temperature in three-dimensional integrated circuits," in *Design Automation Conference, 2006 43rd ACM/IEEE*, 2006, pp. 997–1002.

[132] L. Zhou, C. Wakayama, and C.-J. R. Shi, "CASCADE: A Standard Supercell Design Methodology With Congestion-Driven Placement for Three-Dimensional Interconnect-Heavy Very Large-Scale Integrated Circuits," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 26, no. 7, pp. 1270–1282, 2007.

[133] C.-T. Lin, D.-M. Kwai, Y.-F. Chou, T.-S. Chen, and W.-C. Wu, "CAD reference flow for 3D via-last integrated circuits," in *Design Automation Conference (ASP-DAC), 2010 15th Asia and South Pacific*, 2010, pp. 187–192.

[134] W. R. Davis, E. C. Oh, A. M. Sule, and P. D. Franzon, "Application Exploration for 3-D Integrated Circuits: TCAM, FIFO, and FFT Case Studies," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 4, pp. 496–506, 2009.

[135] J. Cong, A. Jagannathan, Y. Ma, G. Reinman, Y. Zhang, and J. Wei, "An automated design flow for 3D microarchitecture evaluation," in *Proceedings of the 2006 Asia and South Pacific Design Automation Conference*, 2006, pp. 384–389.

[136] C. Mineo, "Clock Tree Insertion and Verification for 3D Integrated Circuits," Master Thesis, North Caroline State University, 2005.

[137] Y.-J. Lee and S. K. Lim, "Timing Analysis and Optimization for Many-Tier 3D ICs," in *Semiconductor Research Corporation (SRC) Technical Conference (TECHCON)*, 2010.

[138] Y. J. Wu, D. Houzet, and S. Huet, "A Programming Model and a NoC-Based Architecture for Streaming Applications," in *Digital System Design: Architectures, Methods and Tools (DSD), 2010 13th Euromicro Conference on*, 2010, pp. 393–397.

[139] U. Y. Ogras, R. Marculescu, H. G. Lee, P. Choudhary, D. Marculescu, M. Kaufman, and P. Nelson, "Challenges and Promising Results in NoC Prototyping Using FPGAs," *Micro, IEEE*, vol. 27, no. 5, pp. 86–95, 2007.

[140] S. V Tota, M. R. Casu, M. R. Roch, L. Macchiarulo, and M. Zamboni, "A Case Study for NoC-Based Homogeneous MPSoC Architectures," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 3, pp. 384–388, 2009.

[141] S. Evain, J.-P. Diguet, and D. Houzet, "uspider: a CAD tool for efficient NoC design," in *Norchip Conference, 2004. Proceedings*, 2004, pp. 218–221.

[142] O. Hammami and M. H. Jabbar, "NoC-based MPSoC Design and Implementation on FPGA: DCT Application," in *Quality Electronic Design (ASQED), 2012 4th Asia Symposium on*, 2012, pp. 304–312.

[143] B. S. Feero and P. P. Pande, "Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation," *Computers, IEEE Transactions on*, vol. 58, no. 1, pp. 32–45, 2009.

[144] V. De Paulo and C. Ababei, "3D Network-on-Chip Architectures Using Homogeneous Meshes and Heterogeneous Floorplans," *International Journal of Reconfigurable Computing*, vol. 2010, pp. 1–12, 2010.

[145] D. Park, S. Eachempati, R. Das, A. K. Mishra, Y. Xie, N. Vijaykrishnan, and C. R. Das, "MIRA: A Multi-layered On-Chip Interconnect Router Architecture," in *Computer Architecture, 2008. ISCA '08. 35th International Symposium on*, 2008, pp. 251–261.

[146] V. Pavlidis and E. Friedman, "3-D Topologies for Networks-on-Chip," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 15, no. 10, pp. 1081–1090, Oct. 2007.

[147] F. Darve, A. Sheibanyrad, P. Vivet, and F. Petrot, "Physical Implementation of an Asynchronous 3D-NoC Router Using Serial Vertical Links," in *VLSI (ISVLSI), 2011 IEEE Computer Society Annual Symposium on*, 2011, pp. 25–30.

[148] P. Vivet, D. Dutoit, Y. Thonnart, and F. Clermidy, "3D NoC using through silicon Via: An asynchronous implementation," in *VLSI and System-on-Chip (VLSI-SoC), 2011 IEEE/IFIP 19th International Conference on*, 2011, pp. 232–237.

[149] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. S. Yousif, and C. R. Das, "A novel dimensionally-decomposed router for on-chip communication in 3D architectures," in *Proceedings of the 34th annual international symposium on Computer architecture*, 2007, pp. 138–149.

[150] D. L. Lewis, S. Yalamanchili, and H.-H. S. Lee, "High Performance Non-blocking Switch Design in 3D Die-Stacking Technology," in *VLSI, 2009. ISVLSI '09. IEEE Computer Society Annual Symposium on*, 2009, pp. 25–30.

[151] C. Liu, L. Zhang, Y. Han, and X. Li, "Vertical interconnects squeezing in symmetric 3D mesh Network-on-Chip," in *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, 2011, pp. 357–362.

[152] Y.-J. Lee and S. K. Lim, "Timing analysis and optimization for 3D stacked multi-core microprocessors," in *3D Systems Integration Conference (3DIC), 2010 IEEE International*, 2010, pp. 1–7.

[153] M. Pathak, Y.-J. Lee, T. Moon, and S. K. Lim, "Through-silicon-via management during 3D physical design: When to add and how many?," in *Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on*, 2010, pp. 387–394.

[154] T. Thorolfsson, G. Luo, J. Cong, and P. D. Franzon, "Logic-on-logic 3D integration and placement," in *3D Systems Integration Conference (3DIC), 2010 IEEE International*, 2010, pp. 1–4.

[155] K. Puttaswamy and G. H. Loh, "3D-Integrated SRAM Components for High-Performance Microprocessors," *Computers, IEEE Transactions on*, vol. 58, no. 10, pp. 1369–1381, 2009.

[156] R. S. Patti, "Homogeneous 3D Integration," in *Three Dimensional System Integration: IC Stacking Process and Design*, A. Papanikolaou, D. Soudris, and R. Radojcic, Eds. Springer US, 2011, pp. 51–71.

[157] S. Microelectronic, "45 nm ST Microelectronic standard library." [Online]. Available: http://cmp.imag.fr/products/ic/?p=STCMOS040.

[158] R. S. Anigundi, H. Sun, J.-Q. Lu, K. Rose, and T. Zhang, "Architecture design exploration of three-dimensional (3D) integrated DRAM," in *Quality of Electronic Design, 2009. ISQED 2009. Quality Electronic Design*, 2009, pp. 86–90.

[159] T. C. Xu, A. W. Yin, P. Liljeberg, and H. Tenhunen, "A study of 3D Network-on-Chip design for data parallel H.264 coding," in *NORCHIP, 2009*, 2009, pp. 1–6.

[160] C. Decayeux and D. Seme, "3D hexagonal network: modeling, topological properties, addressing scheme, and optimal routing algorithm," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 16, no. 9, pp. 875–884, 2005.

[161] J. Zhao, S. Madduri, R. Vadlamani, W. Burleson, and R. Tessier, "A Dedicated Monitoring Infrastructure for Multicore Processors," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 6, pp. 1011–1022, 2011.

[162] W.-K. Mak and C. Chu, "Rethinking the Wirelength Benefit of 3-D Integration," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. PP, no. 99, pp. 1–5, 2011.

[163] T. Thorolfsson, N. Moezzi-Madani, and P. D. Franzon, "Reconfigurable five-layer three-dimensional integrated memory-on-logic synthetic aperture radar processor," *Computers & Digital Techniques, IET*, vol. 5, no. 3, pp. 198–204, 2011.

[164] H. Hua, "Design and Verification Methodology for Complex Three-Dimensional Digital Integrated Circuit," 2006.

[165] K. Yang, D. H. Kim, and S. K. Lim, "Design quality tradeoff studies for 3D ICs built with nano-scale TSVs and devices," in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, 2012, pp. 740–746.

[166] M. Pathak and S. K. Lim, "Performance and Thermal-Aware Steiner Routing for 3-D Stacked ICs," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 28, no. 9, pp. 1373–1386, 2009.

[167] J. Cong and G. Luo, "A multilevel analytical placement for 3D ICs," in *Design Automation Conference, 2009. ASP-DAC 2009. Asia and South Pacific*, 2009, pp. 361–366.

[168] J. Cong and G. Luo, "Advances and Challenges in 3D Physical Design," *IPSJ Transactions on System LSI Design Methodology*, vol. 3, pp. 2–18, 2010.

[169] S. K. Lim, "Tsv-aware 3D physical design tool needs for faster mainstream acceptance of 3D ics," *ACM DAC Knowledge Center (dac.com)*, 2010.

[170] J. L. Ayala, A. Sridhar, and D. Cuesta, "Thermal modeling and analysis of 3D multi-processor chips," *Integration, the VLSI Journal*, vol. 43, no. 4, pp. 327–341, Sep. 2010.

[171] O. Hammami, A. Mzah, M. H. Jabbar, and D. Houzet, "3D IC Implementation for MPSOC Architectures: Mesh and Butterfly Based NoC," in *Quality Electronic Design (ASQED), 2012 4th Asia Symposium on*, 2012, pp. 169–173.

[172] D. H. Kim, K. Athikulwongse, M. B. Healy, M. Hossain, M. Jung, I. Khorosh, G. Kumar, Y.-J. Lee, D. L. Lewis, T.-W. Lin, C. Liu, S. Panth, M. Pathak, M. Ren, G. Shen, T. Song, D. H. Woo, X. Zhao, J. Kim, H. Choi, G. H. Loh, H.-H. S. Lee, and S. K. Lim, "3D-MAPS: 3D Massively parallel processor with stacked memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, 2012, pp. 188–190.

[173] W. Wolf, A. A. Jerraya, and G. Martin, "Multiprocessor System-on-Chip (MPSoC) Technology," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 27, no. 10, pp. 1701–1713, 2008.

[174] P. P. Pande, C. Grecu, A. Ivanov, R. Saleh, and G. De Micheli, "Design, synthesis, and test of networks on chips," *Design & Test of Computers, IEEE*, vol. 22, no. 5, pp. 404–413, 2005.

[175] U. Y. Ogras, R. Marculescu, D. Marculescu, and E. G. Jung, "Design and Management of Voltage-Frequency Island Partitioned Networks-on-Chip," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 3, pp. 330–341, 2009.

[176] M. Krstic, E. Grass, F. K. Gurkaynak, and P. Vivet, "Globally Asynchronous, Locally Synchronous Circuits: Overview and Outlook," *Design & Test of Computers, IEEE*, vol. 24, no. 5, pp. 430–441, 2007.

[177] D. H. Woo, N. H. Seong, and H.-H. S. Lee, "Heterogeneous die stacking of SRAM row cache and 3-D DRAM: An empirical design evaluation," in *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on*, 2011, pp. 1–4.

[178] H. Sun, J. Liu, R. S. Anigundi, N. Zheng, J.-Q. Lu, K. Rose, and T. Zhang, "Design of 3D DRAM and Its Application in 3D Integrated Multi-Core Computing Systems," *Design & Test of Computers, IEEE*, vol. 26, no. 5, pp. 36–47, 2009.

[179] V. S. Nandakumar and M. Marek-Sadowska, "A Low Energy Network-on-Chip Fabric for 3-D Multi-Core Architectures," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 2, no. 2, pp. 266–277, 2012.

[180] K.-S. Chong, K.-L. Chang, B.-H. Gwee, and J. S. Chang, "Synchronous-Logic and Globally-Asynchronous-Locally-Synchronous (GALS) Acoustic Digital Signal Processors," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 3, pp. 769–780, 2012.

[181] L. A. Plana, S. B. Furber, S. Temple, M. Khan, Y. Shi, J. Wu, and S. Yang, "A GALS Infrastructure for a Massively Parallel Multiprocessor," *Design & Test of Computers, IEEE*, vol. 24, no. 5, pp. 454–463, 2007.

[182] A. R. Marschner, S. Craven, and P. Athanas, "A Sandbox For Exploring The Openfire Processor," in *ERSA'07*, 2007, pp. 248–251.

[183] T. Kranenburg, "Design of a Portable and Customizable Microprocessor for Rapid System Prototyping," Master Thesis, Delft University, 2009.

[184] S. Craven, C. Patterson, and P. Athanas, "Configurable Soft Processor Arrays Using the OpenFire Processor," in *MAPLD 05*, 2005.

[185] A. R. Marschner, "An FPGA-based Target Acquisition System," Master Thesis, Virginia Polytechnic Institute and State University, 2007.

[186] Z. Yu and B. M. Baas, "High Performance, Energy Efficiency, and Scalability With GALS Chip Multiprocessors," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 17, no. 1, pp. 66–79, 2009.

[187] S. R. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, "An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 1, pp. 29–41, 2008.

[188] M. Grange, A. Y. Weldezion, D. Pamunuwa, R. Weerasekera, Z. Lu, A. Jantsch, and D. Shippen, "Physical mapping and performance study of a multi-clock 3-Dimensional Network-on-Chip mesh," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, 2009, pp. 1–7.

[189] V. F. Pavlidis, I. Savidis, and E. Friedman, "Clock Distribution Networks in 3-D Integrated Systems," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 12, pp. 2256–2266, 2011.

[190] X. Zhao, S. Mukhopadhyay, and S. K. Lim, "Variation-tolerant and low-power clock network design for 3D ICs," in *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st*, 2011, pp. 2007–2014.

[191] C.-L. Lung, Y.-S. Su, S.-H. Huang, Y. Shi, and S.-C. Chang, "Fault-tolerant 3D clock network," in *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, 2011, pp. 645–651.

[192] X. Zhao and S. K. Lim, "Through-silicon-via-induced obstacle-aware clock tree synthesis for 3D ICs," in *Design Automation Conference (ASP-DAC), 2012 17th Asia and South Pacific*, 2012, pp. 347–352.

[193] J.-S. Yang, J. Pak, X. Zhao, S. K. Lim, and D. Z. Pan, "Robust Clock Tree Synthesis with timing yield optimization for 3D-ICs," in *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, 2011, pp. 621–626.

[194] X. Zhao, D. L. Lewis, H.-H. S. Lee, and S. K. Lim, "Low-Power Clock Tree Design for Pre-Bond Testing of 3-D Stacked ICs," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 30, no. 5, pp. 732–745, 2011.

[195] M. B. Healy, "Performance and Temperature Aware Floorplanning Optimization for 2D and 3D Microarchitectures," Master Thesis, Georgia Institute of Technology, 2006.

[196] T. Thorolfsson, "Three-Dimensional Integration of Synthetic Aperture Radar Processors," PhD Thesis, North Carolina State University, 2011.

[197] P. Salihundam, S. Jain, T. Jacob, S. Kumar, V. Erraguntla, Y. Hoskote, S. Vangal, G. Ruhl, and N. Borkar, "A 2 Tb/s 6 x 4 Mesh Network for a Single-Chip Cloud Computer With DVFS in 45 nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 4, pp. 757–766, 2011.

[198] A. Ivankovic, G. Van der Plas, V. Moroz, M. Choi, V. Cherman, A. Mercha, P. Marchal, M. Gonzalez, G. Eneman, W. Zhang, T. Buisson, M. Detalle, A. L. Manna, D. Verkest, G. Beyer, E. Beyne, B. Vandevelde, I. De Wolf, and D. Vandepitte, "Analysis of microbump induced stress effects in 3D stacked IC technologies," in *3D Systems Integration Conference (3DIC), 2011 IEEE International*, 2012, pp. 1–5.

[199] S. Priyadarshi, J. Hu, W. H. Choi, S. Melamed, X. Chen, W. R. Davis, and P. D. Franzon, "Pathfinder 3D: A flow for system-level design space exploration," in *3D Systems Integration Conference (3DIC), 2011 IEEE International*, 2012, pp. 1–8.

[200] D. Milojevic, T. E. Carlson, K. Croes, R. Radojcic, D. F. Ragett, D. Seynhaeve, F. Angiolini, G. Van der Plas, and P. Marchal, "Automated Pathfinding tool chain for 3D-stacked integrated circuits: Practical case study," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, 2009, pp. 1–6.

[201] D. Milojevic, R. Radojcic, R. Carpenter, and P. Marchal, "Pathfinding: A design methodology for fast exploration and optimisation of 3D-stacked integrated circuits," in *System-on-Chip, 2009. SOC 2009. International Symposium on*, 2009, pp. 118–123.

[202] D. Diamantopoulos, K. Siozios, and D. Soudris, "Framework for performing rapid evaluation of 3D SoCs," *Electronics Letters*, vol. 48, no. 12, pp. 679–681, 2012.

[203] D. Diamantopoulos, K. Siozios, D. Bekiaris, and D. Soudris, “A novel methodology for architecture-level exploration of 3D SoCs,” in *Design & Technology of Integrated Systems in Nanoscale Era (DTIS), 2011 6th International Conference on*, 2011, pp. 1–6.

[204] N. A. V Doan, F. Robert, Y. De Smety, and D. Milojevic, “MCDA-based methodology for efficient 3D-design space exploration and decision,” in *System on Chip (SoC), 2010 International Symposium on*, 2010, pp. 76–83.

[205] A. Richard, D. Milojevic, F. Robert, A. Bartzas, A. Papanikolaou, K. Siozios, and D. Soudris, “Fast Design Space Exploration Environment Applied on NoC’s for 3D-Stacked MPSoC's,” in *Architecture of Computing Systems (ARCS), 2010 23rd International Conference on*, 2010, pp. 1–6.

[206] A. M’zah, O. Hammami, and J. Mouine, “The Impact of EDA Tools in 3D IC Design Space Exploration: A Case Study,” in *DATE 2012 Workshop: 3D Integration - Application, Technology, Design, Automation and Test*, 2012.